New ERC correlation

Monday, August 8, 2016 2:16 PM

Brandon Douglas Jernigan **<brandonjernigan@email.arizona.edu>** Jun 6

to Ryan

Hi Ryan, I'm back from my vacation to Sedona, Grand Canyon, and Payson.

I finally got some data that correlated with our results. On Nathan's ERC webserver (http://csb.pitt.edu/erc_analysis/) if I put in a selection of the genes we are looking at into group analysis, those numbers compare well with ours (scatterplot is attached, r = 0.679). The ERC matrix I was using when there wasn't good correlation was the one from supplementary material (http://www.csb.pitt.edu/faculty/clark/data.html) in the paper we've been trying to recreate the results of (*Evolutionary rate covariation reveals shared functionality and coexpression of genes*). It is identical to the yeast matrix that can be downloaded from the ERC webserver in the methods section (http://csb.pitt.edu/erc_analysis/Methods.php), so it's confusing that using the webserver yields different results than downloading the dataset it's supposedly based on.

I discovered this was the case when I tried to recreate figure 4 from the ERC paper above using values from the ERC matrix in the supplementary material and the values didn't match up, but the webserver results did. Everything else seemed to check out in my pipeline. I would like to get all of the data from the webserver, but my browser starts crashing after matrices larger than 100 x 100 genes, which means I'd need to load about 2,000 of those to construct the full ~ 5,000 x 5,000 matrix, so maybe there's a better way to get that data.

Also, it seems like I'm not set up properly in the payroll system or something, when I look for the timecard in UAccess, it isn't there anymore.

Attachments area

Gutenkunst, Ryan N - (rgutenk) **<rgutenk@email.arizona.edu>** Jun 11

to me

Hi Brandon,
Sorry for the very slow reply. I've had little good email access while traveling.
Great that you're seeing good ERC results finally. It is worrying that the matrix Nathan published doesn't line up with the web server. Might bet is that there's a simple mistake in that big matrix. Let's nail down a few simple examples and let Nathan know about it.
In the mean time, I think we can move on to getting domain statistics.
Let me check in about payroll.
Best,
Ryan

From <https://mail.google.com/mail/u/1/#search/rgutenk%40email.arizona.edu/15527b50f61e96ee>

06-13-2016 Domain stats

Monday, August 8, 2016 2:14 PM

Brandon Douglas Jernigan <brandonjernigan@email.arizona.edu> Jun 13

to Ryan

Hi Ryan, it looks like the differences between the webserver data and the downloaded matrices was due to a script of mine I was using to translate the gene names. It was collapsing any empty columns and causing all the values to get thrown off. Once that's fixed the webserver values and matrix values match up.
I've been corresponding with someone from hpc-consult trying to figure out how why I'm having so many issues running things successfully on the hpc. It appears one of the issues was that windows newline characters had crept in from editing the PBS scripts on my desktop. There were also lines that seemed to use weird characters, because when I re-typed them exactly, they suddenly worked. I'm still trying to make sure I can run everything I need to on the hpc, so hopefully most of the issues can get worked out this week.
Brian and I are going to meet this Wednesday to go over getting domain statistics. I'm going to re-read the "Selection on network dynamics drives differential rates of
protein domain evolution" paper, but is there anything else you think I should look at? And is there anything specific you want me to start looking into in the mean time?
This past weekend I also attended the software carpentry workshop. I'm feeling much more comfortable with using the command line to do things. They also showed me how to use the cyverse (formerly iplant) cloud computing, and it actually looks like it will be useful. I can have 4 cpus, 32 gb ram, 240 gb drive running on an instance, and I have full control over the system (Linux) to install or change anything. So it looks like it might be a useful resource, in addition to running things on the hpc.
I've also been streamlining the ERC workflow, now everything is more flexible and can be executed with one shell script. Going to see if how it does on the hpc once I get all of that figured out.
I've commited my changes, but it looks like there isn't very much space left on the bitbucket, so I'm going to try to leave out as many unnecessary files as I can.

Gutenkunst, Ryan N - (rgutenk) <rgutenk@email.arizona.edu> Jun 13

to me

Hi Brandon,
Sounds like great progress. In the meantime before meeting with Brian, you can look into some of the databases that annotate domains. Good starting points would be UniProt, Interpro, and PFam.
As for Bitbucket, I'm surprised you're running out of space. Generally we want to store analysis scripts (which should be small) and final results (which should also be mostly small). Big intermediate or binary files shouldn't be stored, nor should anything that can be generated by a script. We can sort through it when I get back.
cyverse does sound like a great resource. I'm glad you're learning it. You should tell me more about it when I get back!
Best,
Ryan

From <https://mail.google.com/mail/u/1/#search/rgutenk%40email.arizona.edu/1554c2fff51550aa>

I2D

Monday, August 8, 2016 1:25 PM

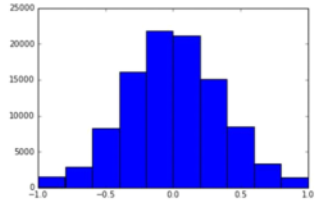
Brandon Douglas Jernigan <brandonjernigan@email.arizona.edu> Jun 27

to Ryan

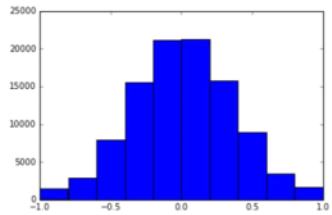
So I was able to get the ERC values from interacting proteins using data from I2D. The values are near -1 and 1 as expected. I've applied the annotations from pfam now, and there are more domains in that data. Looking at the values just based on linker and domain regions of the protein, there doesn't seem to be a very strong signal. Everything is pretty much the same with a mean near 0. Both of those results are in the word document attached.
I had some issues with ID conversions, it looks like some of the conversion tables include a version number (like NP_015151.1 vs NP_015151). So I'm not sure if those sequences are then supposed to be identical or not. It's not really clear yet, but I'll try to find out.
I tried to get results for Brian's gene annotations, but it looks like almost none of our species match up (The only one in common is cerevisiae), I also tried applying his annotations to my sequences, but the lengths are different, which is strange. I'll talk to Brian about that., but until we either get those other species' genomes or find a way to apply his annotations directly we won't be able to get results.
Brian suggested I use python networkx to look at more network properties of genes in relation to ERC.
So this coming week I'll try to resolve some of the things that aren't working and see what kind of network properties are worth looking into.

From <https://mail.google.com/mail/u/1/#search/rgutenk%40email.arizona.edu/155946b81505a906>

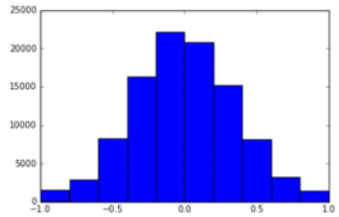
Across all ERC values, mean = -0.000037



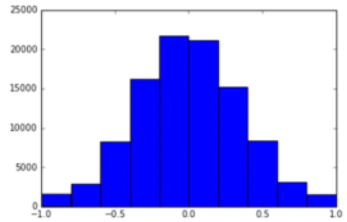
Across all defined domains from pfam, mean = -0.003322



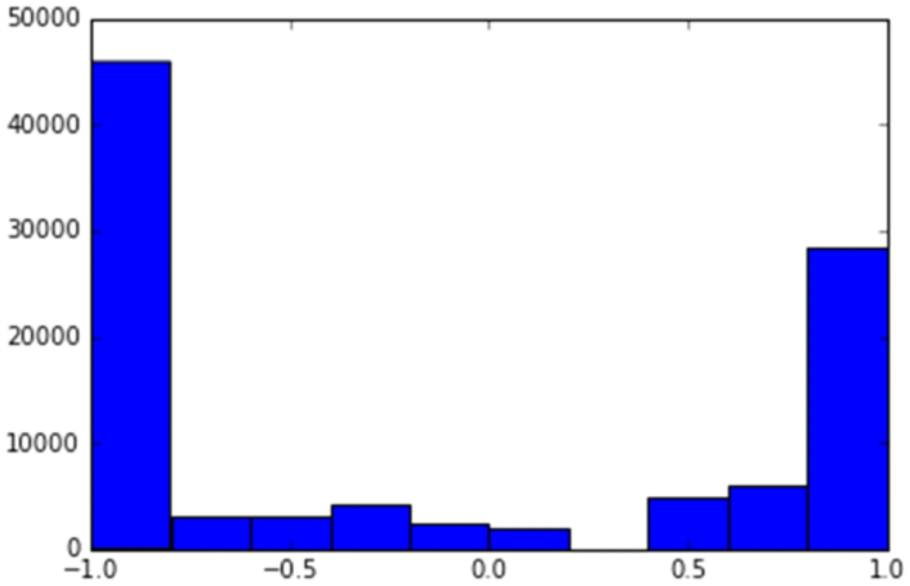
Across all linker regions from pfam, mean = 0.012344



Between linkers and domains, mean = -0.001653



ERC values between interacting genes according to I2D, mean -0.01558



Notepad ++ Notes

Wednesday, July 20, 2016 7:04 PM

use brians domain annotations
collecting erc values for domains
high erc between domains vs linker sequences (proteins with and without multiple domains)
look at interacting domains (have for brians) (pfam interacting domains)(gene-gene interaction network)
I2D (between domains of interacting proteins)(summarize top scores)(compare domain-domain vs protein-protein signal strength)
different ways to look at data
erc for brians domains

python networkx

check between domains on same protein vs domains not on same protein
how to calculate erc p values (uncertainty)
hessian matrix, influence by changing one parameter (diagonal, brian's paper), vs changing 2 (off diagonal)
distribution of p values
access hessians
separate domains, repeated domains

2007 sloppiness paper for hessian

bin erc distribution
degree gives list of nodes interacting
nodes that interact with many interacting partners, vs ones that don't etc.

visualize sections of interaction network with erc coded as color, explore patterns

high values of hessian matrix
Domain-domain interaction based on brian's data
lower pvalues on histogram

geometric average of all unique matrix values for all parameters in all reactions for each protein pair domains interacting within reactions list
combinations of length 2 for finding unique cells for all parameters within matrix in python

sign in geometric mean of influences, is it important (we have erc signs) ask Ryan

Brian, dynamical influence: http://journals.plos.org/plosgenetics/article?id=10.1371%2Fjournal.pgen.1006132

Nathan ERC (Yeast): http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3317153/

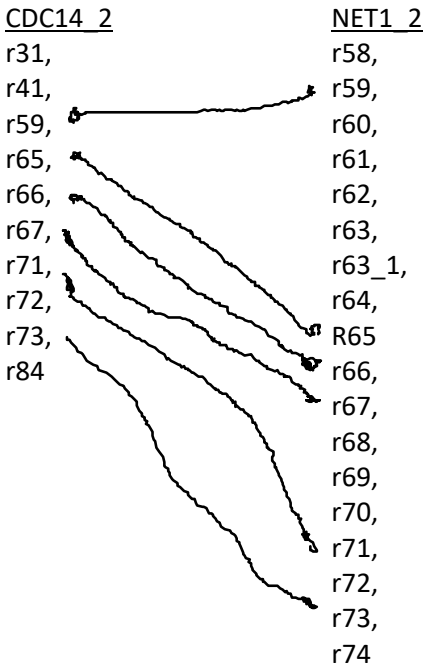
ERC Webserver: http://csb.pitt.edu/erc_analysis/

ERC Vertebrate: http://journals.plos.org/plosgenetics/article?id=10.1371%2Fjournal.pgen.1004967

Pull Pair and Parameters

Tuesday, July 19, 2016 7:56 PM

Vinod2011 data



Both within reactions:
r59,
r65,
r66,
r67,
r71,
r72,
r73,

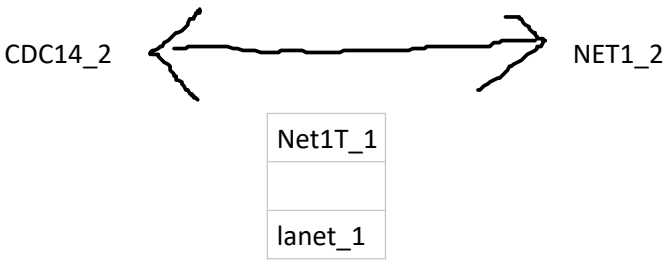
Proteins, the reactions they're in, the parameters associated with reactions

r59,	**r65,**	**r66,**	**r67,**	**r71,**	**r72,**	**r73,**
Jnet_1	Jnet_1			Jnet_1		
Net1T_1	Net1T_1	Net1T_1		Net1T_1	Net1T_1	
kd_1	kd_1			kd_1		
		lanet_1			lanet_1	
			ldnet_1			ldnet_1

r59

Jnet_1
Net1T_1
kd_1

r66



	Jnet_1	Net1T_1	kd_1	lanet_1	ldnet_1
Jnet_1	1.83E-02	?	?	?	?
Net1T_1	-4.74E-03	9.65E-01	?	?	?
kd_1	5.28E-03	-4.95E-02	7.71E-02	?	?
lanet_1	-2.58E-03	5.73E-03	-4.55E-03	2.39E-03	?
ldnet_1	2.74E-03	-5.52E-03	4.39E-03	-2.34E-03	2.32E-03

Options:

- Between parameters in each reaction
- Between all parameters associated with this pair of genes

What if only one parameter associated? Include on-diagonal values?

Can calculate different versions as new columns (include diagonal, without diagonal, geometric absolute mean, arithmetic mean, arithmetic absolute, mark single parameter values[can remove if necessary], number of shared reactions, average values for: dynamical influence)

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n}.$$ Geometric mean of off diagonal with abs vals = 0.005143

[matches value calculated in python script]

Certain parameters show up in several reactions, while others only show up in one. Should those values be weighted to reflect that?

Could count repeated parameters just to keep tabs on how often it occurs.

I think I'll just pull all parameters, not worry about repeats right now and include the variations as new columns as specified above.

Could preserve all values for Brian's data in column with sub-dataframe

Brian:

Given a pair of domains (D1,D2), they share reactions

(D1,D2) -> [R1,R2]
which have parameters
(D1,D2) - > [k1,k2,k3,k4]
which lead to hessian entries
(D1,D2) -> [(k1,k2),(k1,k3),(k1,k4),(k2,k3),(k2,k4),(k3,k4)]
and so

Co-Influence(D1,D2) = f([(k1,k2),(k1,k3),(k1,k4),(k2,k3),(k2,k4),(k3,k4)]) i.e Co-influence(what are we calling this?) is a function of the partial derivatives of shared parameters.

Question #1: What is f?
We have been using geometric mean, but since the partials can be negative, this would be sensitive to the count of signs. If we use an arithmetic mean, large derivatives of a particular sign will dominate. My thinking is that this is actually what we want, but I am not certain of this.

Question #2: What is Co-influence(D1,D2) if D1 and D2 share no reactions.
Missing seems logical, but I could make an argument for 0 as well.

From <https://mail.google.com/mail/u/1/#inbox/155eb62088758d8e>

Ryan:

#1: What is f? That's something we need to explore. I think we mainly care about magnitudes, so I would try a geometric mean of the absolute values, but it's worth exploring other approaches.

#2: If proteins share no reactions, they could still evolve in a compensatory way. So I would calculate co-influence using the union of their parameter sets. For example, protein 1 participates in reactions with parameters k1,k2,k3 and protein 2 participates in reactions with k4,k5,k6. One could calculate co-influence using all nine possible compensatory pairs (k1,k4),(k1,k5)…

From <https://mail.google.com/mail/u/1/#inbox/155eb62088758d8e>

Try network x visualization of domain network vs erc, hessian, degree of separattion, etc

Distribution of hessians between parameters within the same reaction or protein pair

Zoom in on lower pvalues in pvalue histogram for brian's erc (many more bins)

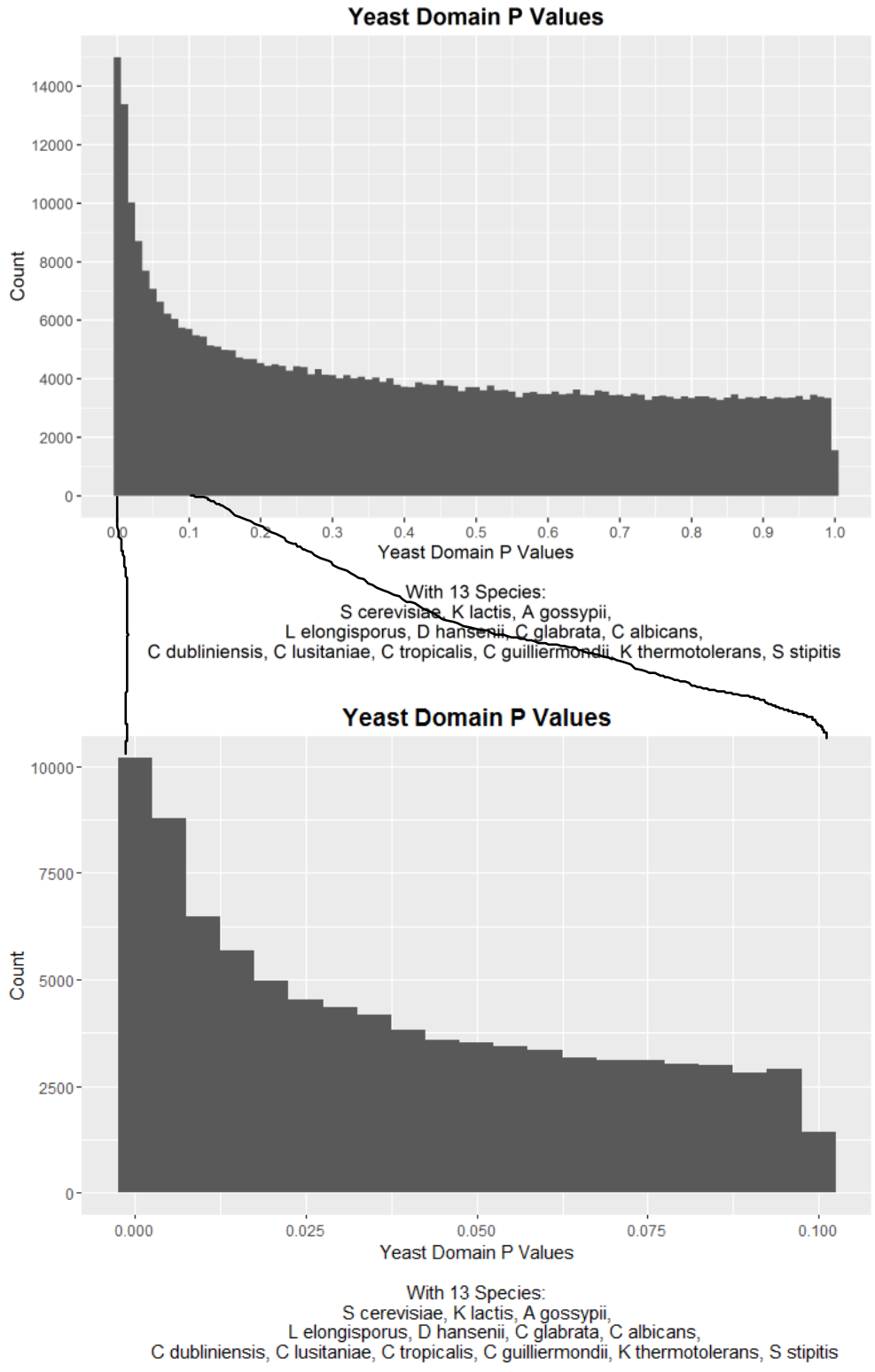
Investigate reason for low p-vals (number of species?)

Interpretation, meaning, next steps, questions.

P value distribution

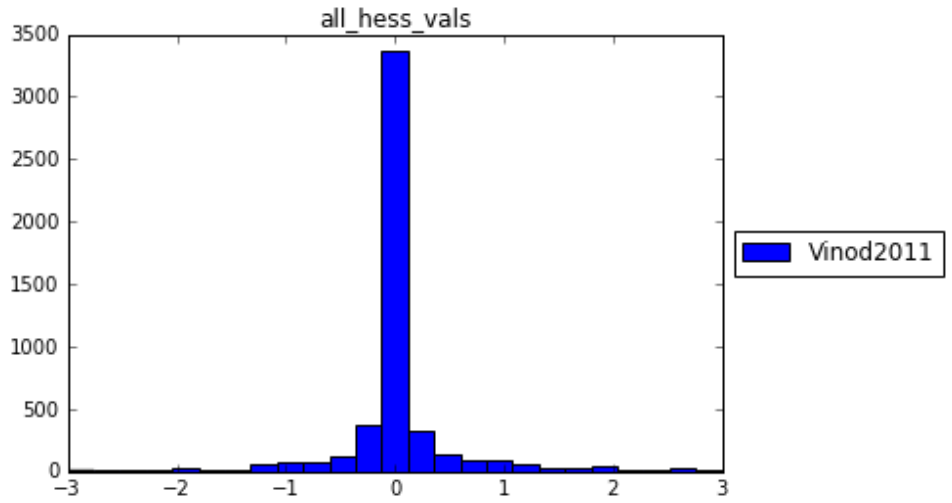
Monday, July 25, 2016 10:51 AM

ERC values calculated from fewer species in the phylogenetic tree tend to have higher p-values, but even trees with large numbers of species have a lot of high p-values. Not a gigantic difference, but there is a difference. P-values below 5% represent ~ 7% of the data (estimated).



Distribution of means in hessian data

Friday, July 22, 2016 3:15 AM

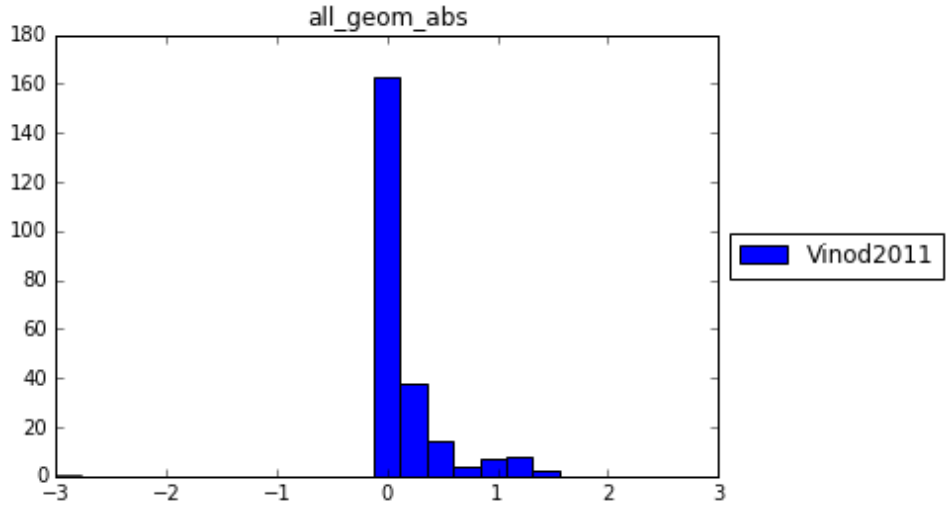


Distribution of all hessian values used in finding means

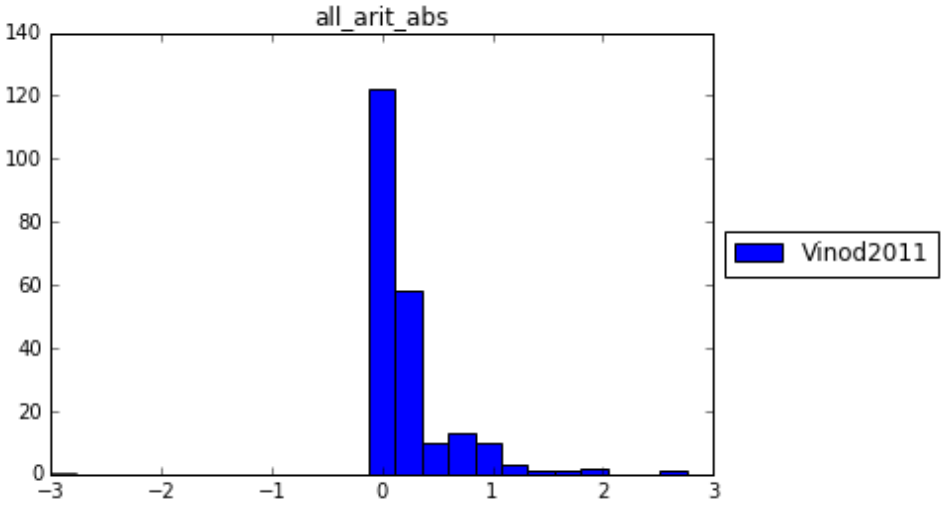
Means:

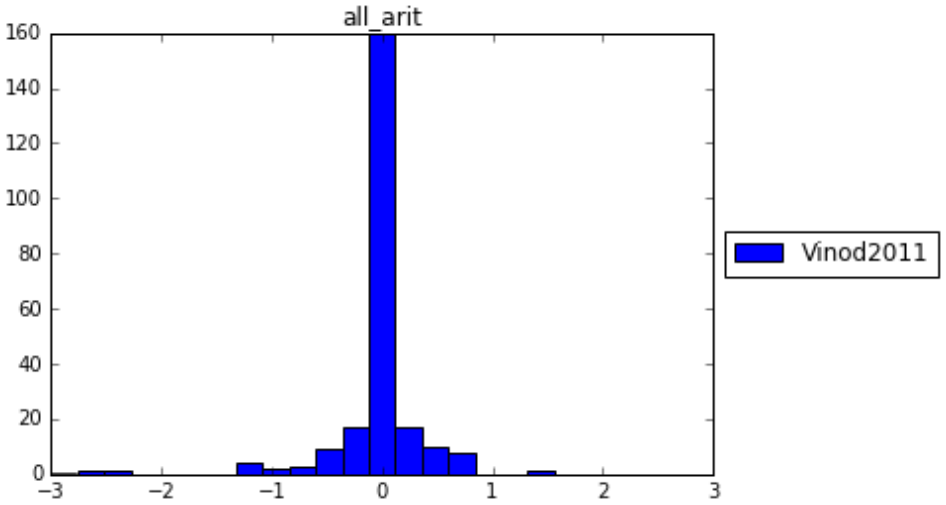
- upper triangular + diagonals of each matrix of domains:
  - upper triangular + diagonals of each domain pair's matrix of reactions (repeats removed):
    - Upper triangular matrix + diagonals of each parameter matrix associated with the reactions (how to handle single parameters?)

--



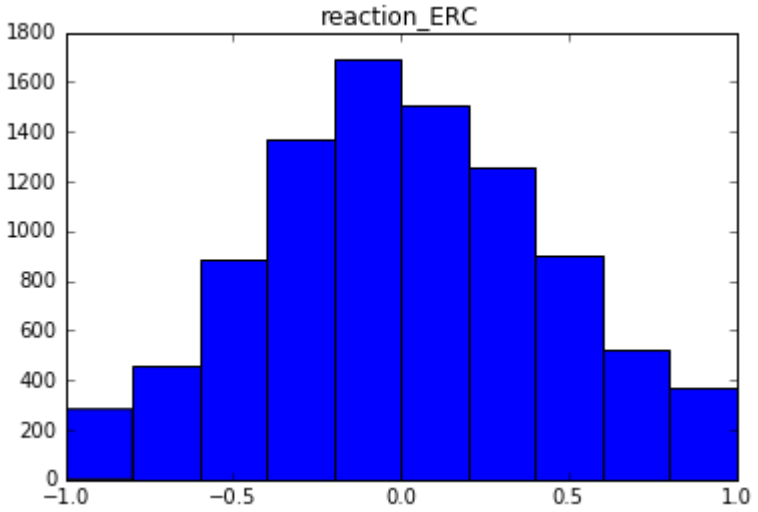
Distribution of geometric absolute mean



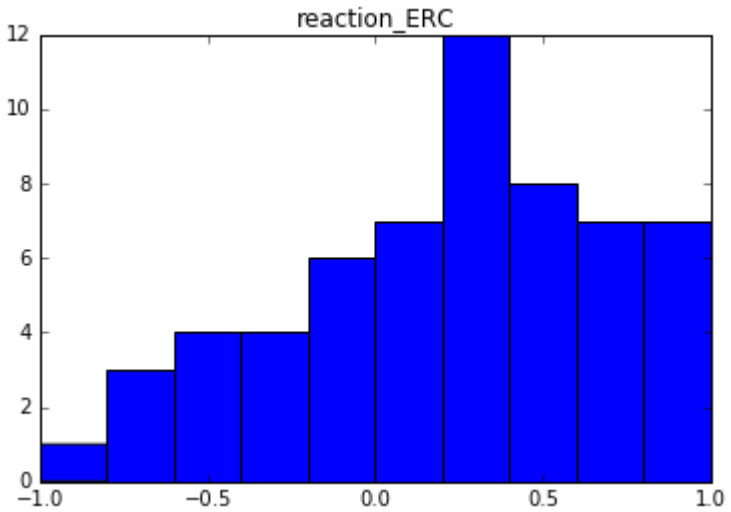


ERC distribution between proteins sharing reactions

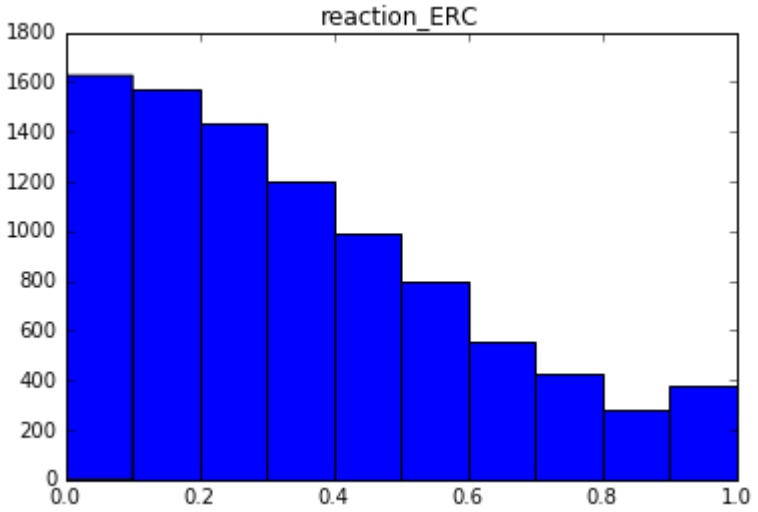
Monday, July 25, 2016 12:46 AM



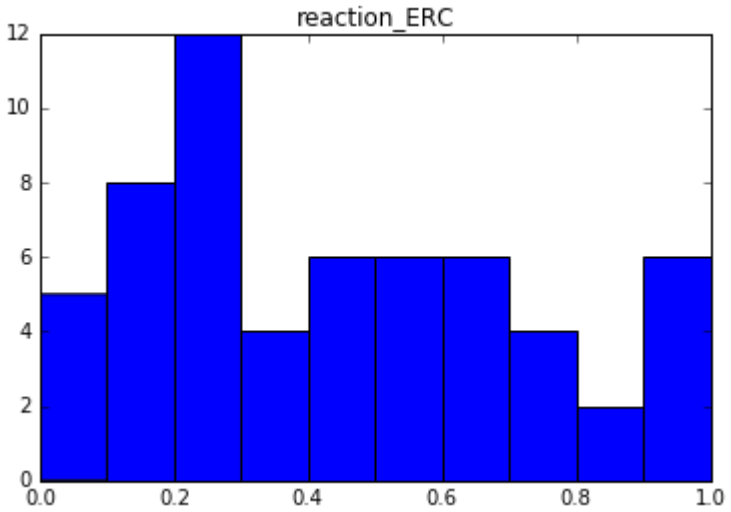
Distribution of all brian's erc values
Mean = 0.00864



Distribution of brians erc values with proteins sharing a reaction
Mean = 0.21917



Distribution of absolute values of all brian's erc values
Mean = 0.35085



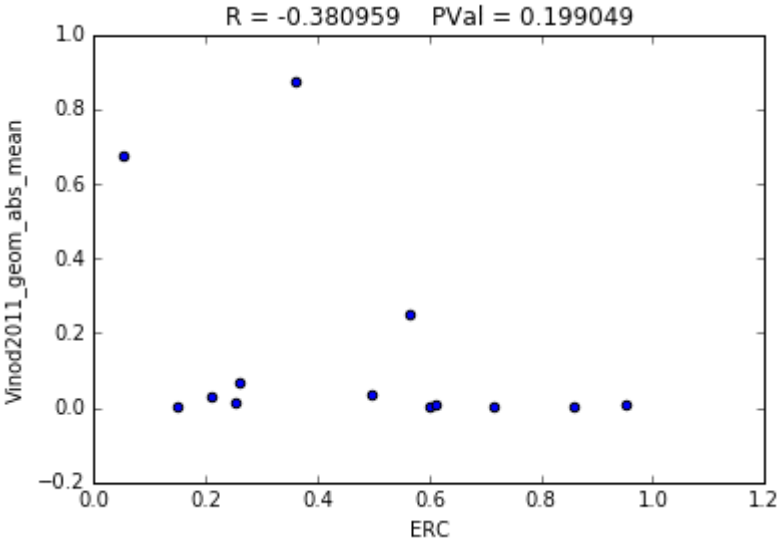
Distribution of absolute values of brians erc values with proteins sharing a reaction
Mean = 0.4384

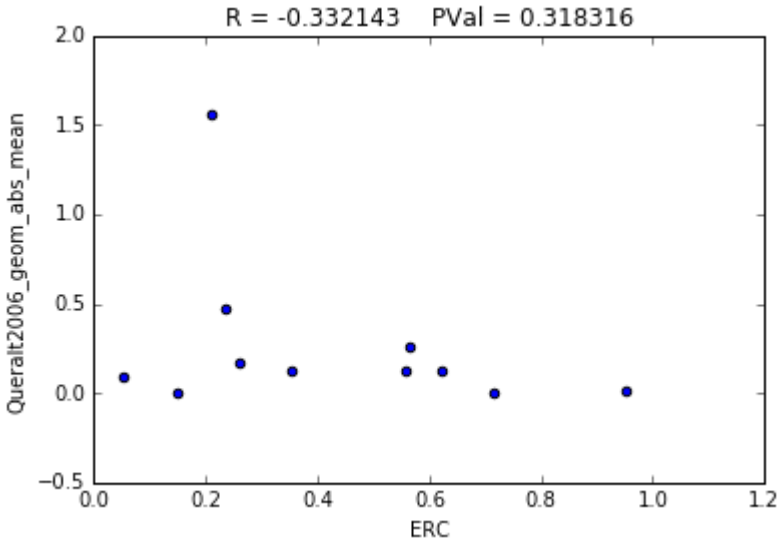
Shared geometric absolute mean of hessian data vs erc

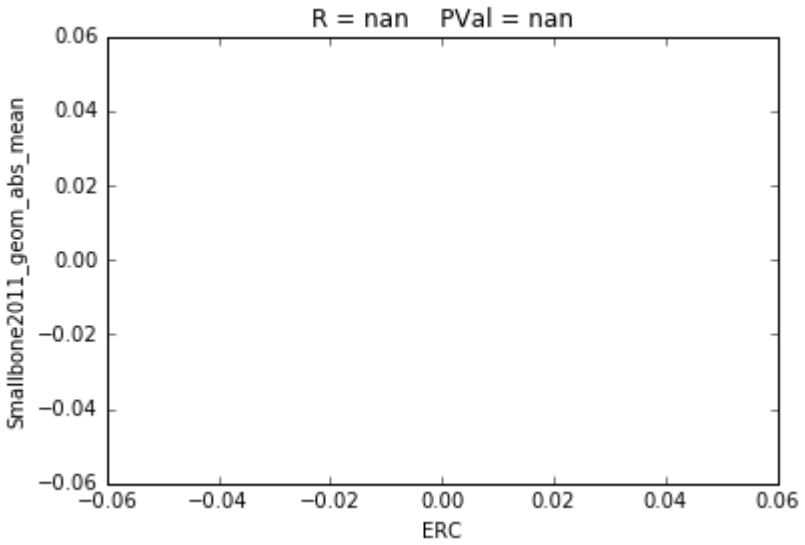
Sunday, July 24, 2016 10:22 PM

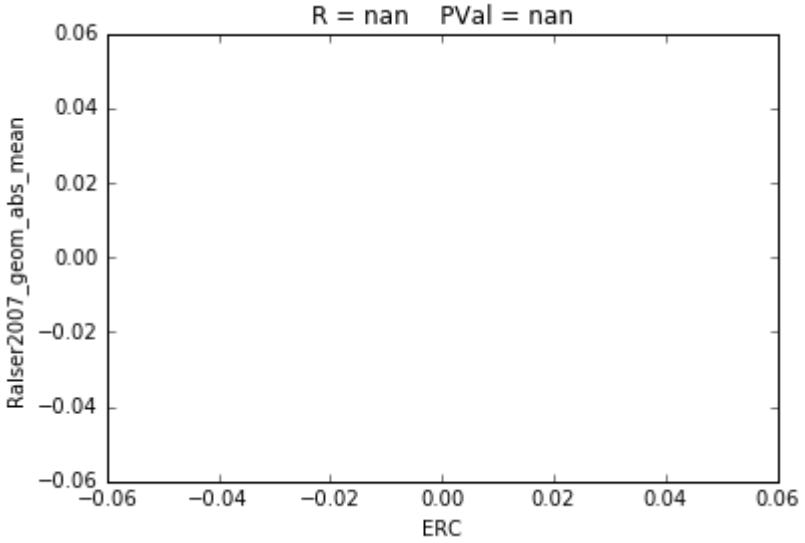
For proteins sharing a reaction, parameters in common between shared reactions,
off diagonal hessian entries

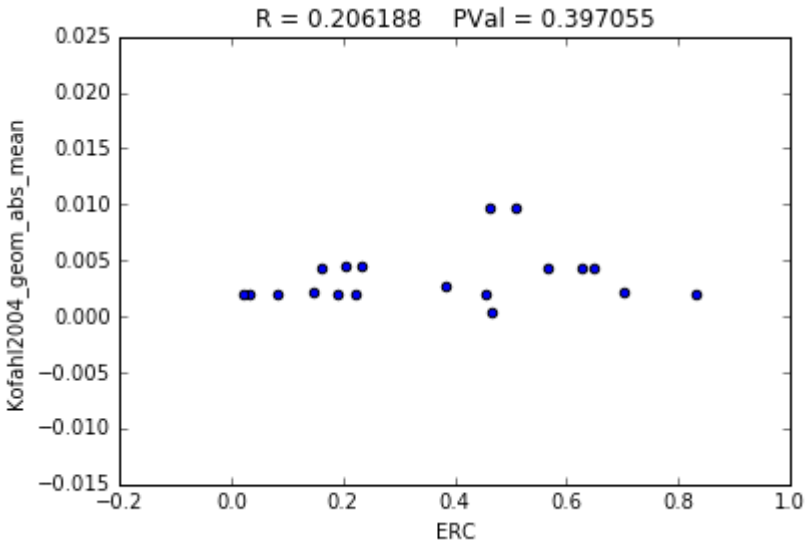
Empty scatter plots due to 0 or 1 (1 leaves only a diagonal entry) hessian parameter.
Many protein pairs did not have overlapping interactions.

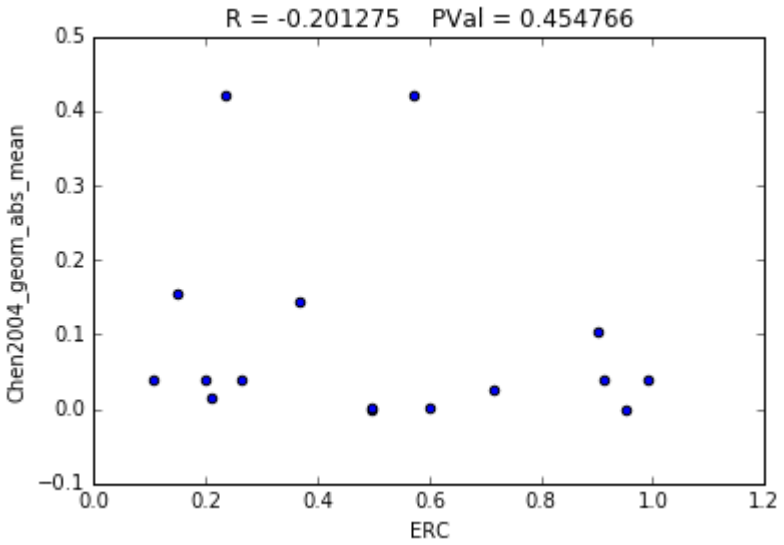





R = nan PVal = nan
Smallbone2011_geom_abs_mean
ERC


R = nan PVal = nan
Ralser2007_geom_abs_mean
ERC


R = 0.206188 PVal = 0.397055
Kofahl2004_geom_abs_mean
ERC

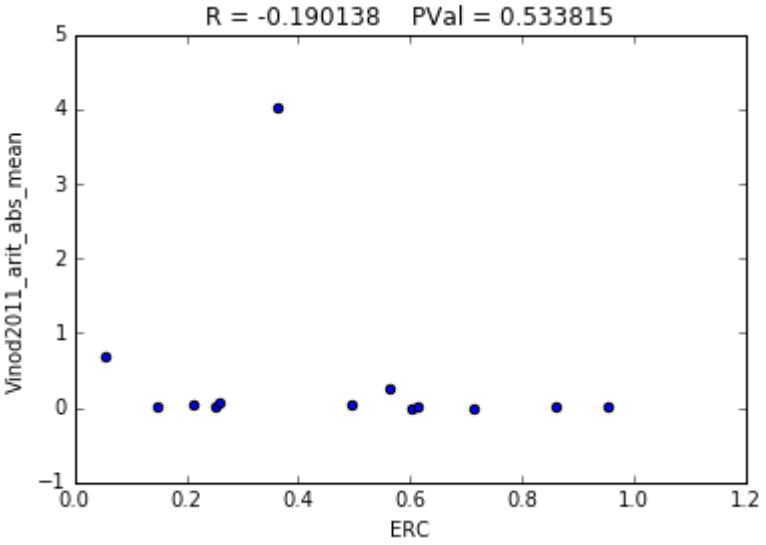

R = -0.201275 PVal = 0.454766
Chen2004_geom_abs_mean
ERC
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
1.2

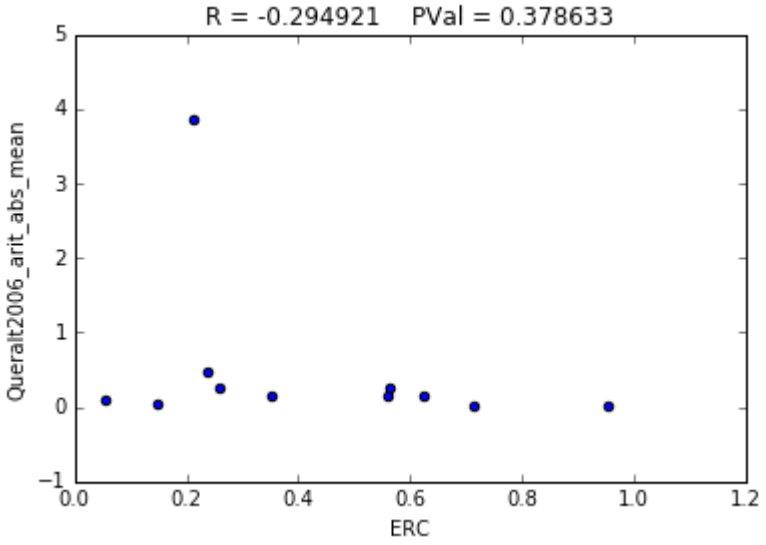
Shared arithmetic absolute mean of hessian data vs erc

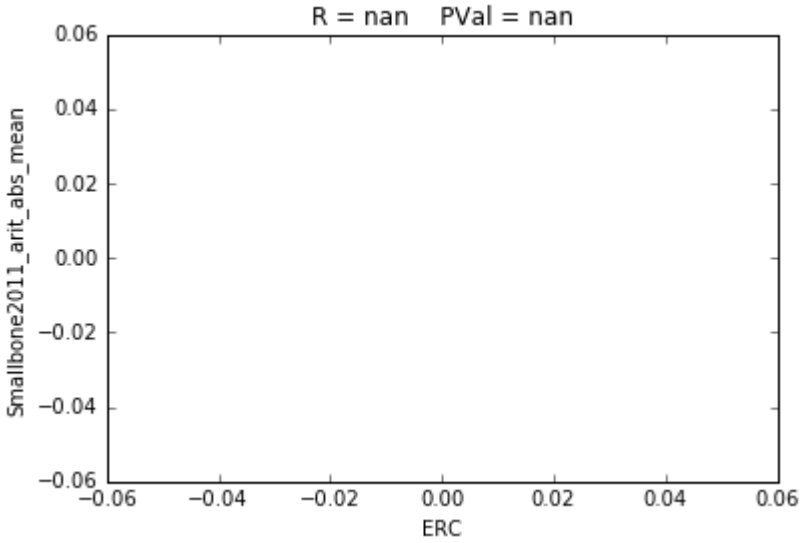
Monday, July 25, 2016 1:13 AM

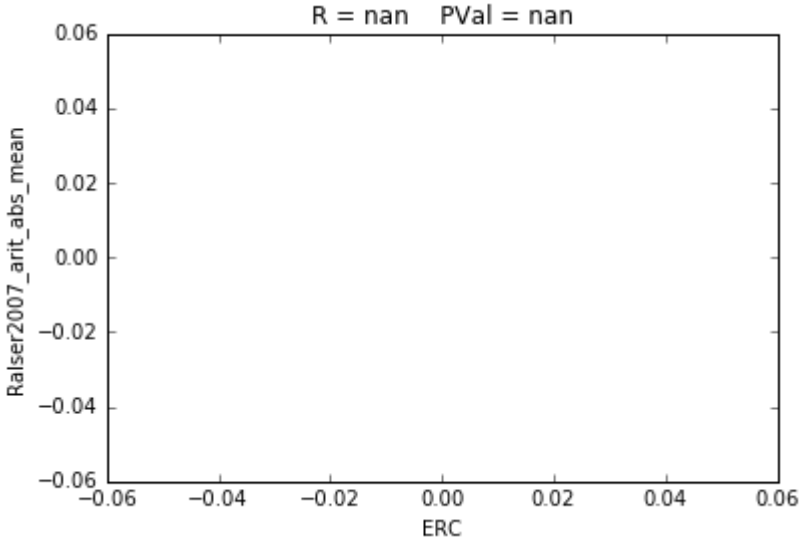
For proteins sharing a reaction, parameters in common between shared reactions,
off diagonal hessian entries

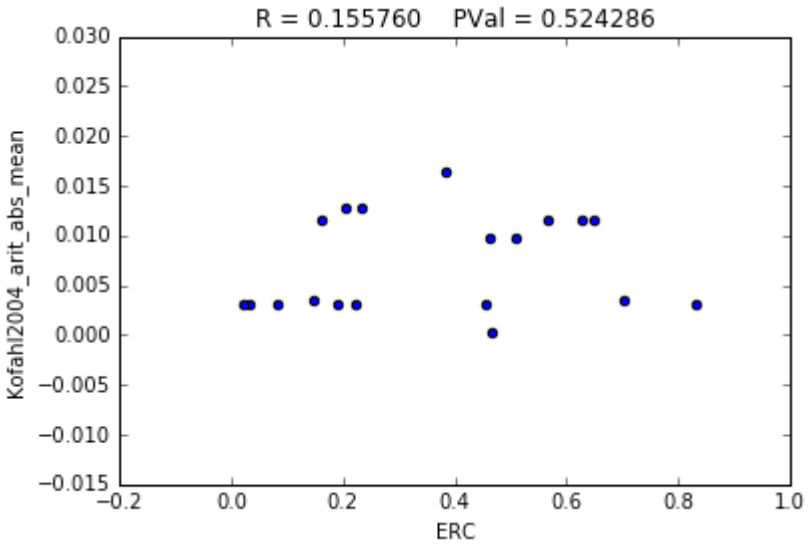
Empty scatter plots due to 0 or 1 (1 leaves only a diagonal entry) hessian parameter.
Many protein pairs did not have overlapping interactions.

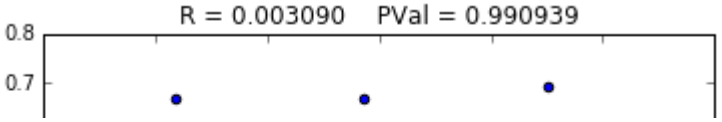


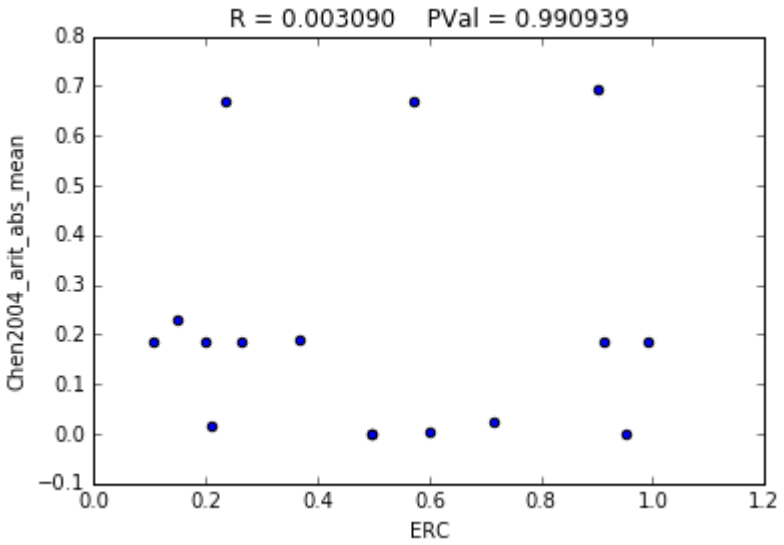



R = nan PVal = nan
Smallbone2011_arit_abs_mean
ERC


R = nan PVal = nan
Ralser2007_arit_abs_mean
ERC


R = 0.155760 PVal = 0.524286
Kofahl2004_arit_abs_mean
ERC


R = 0.003090 PVal = 0.990939

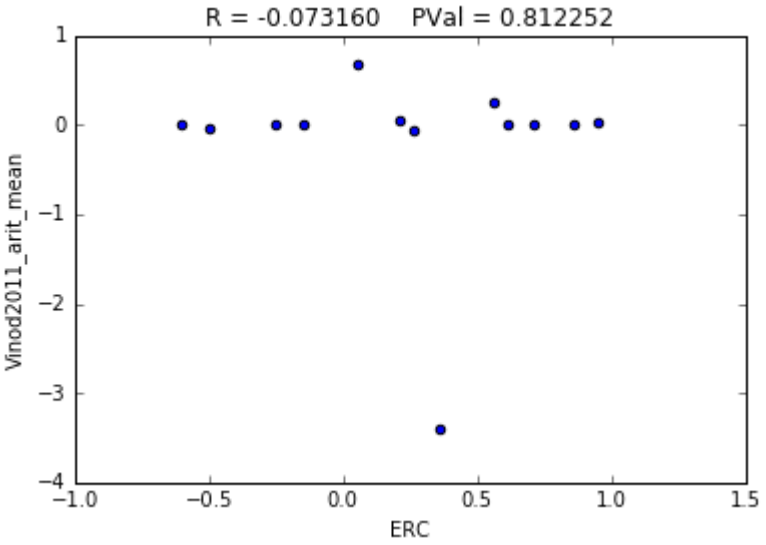

R = 0.003090 PVal = 0.990939
Chen2004_arit_abs_mean
ERC
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
1.2

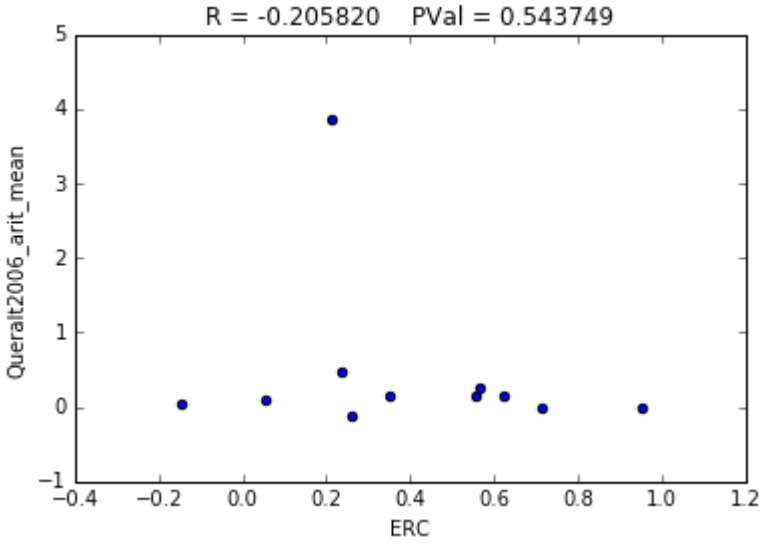
Shared arithmetic mean of hessian data vs erc

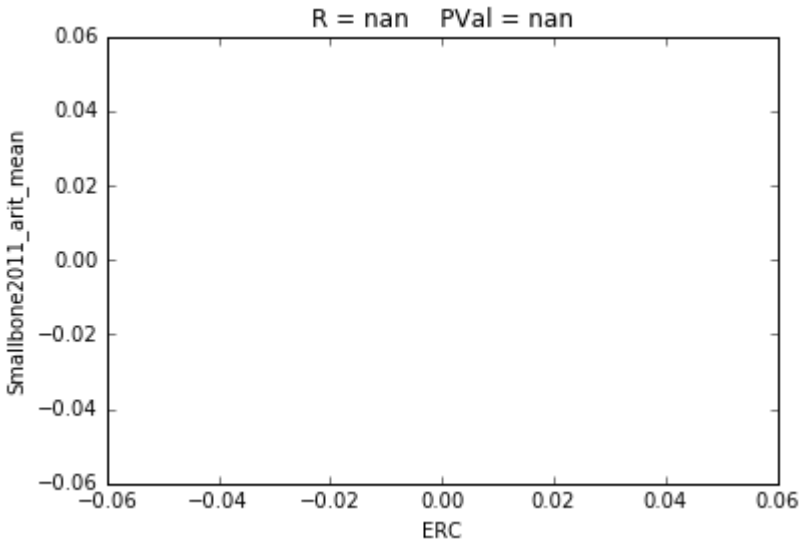
Monday, July 25, 2016 1:15 AM

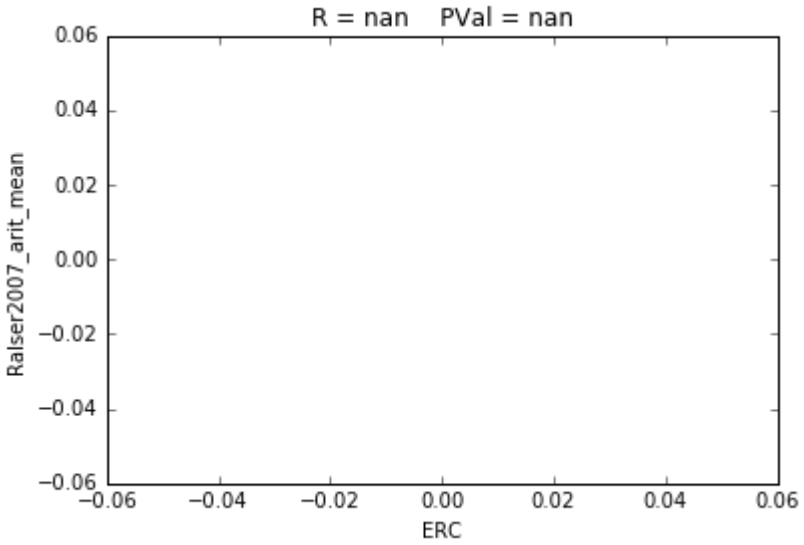
For proteins sharing a reaction, parameters in common between shared reactions, off diagonal hessian entries

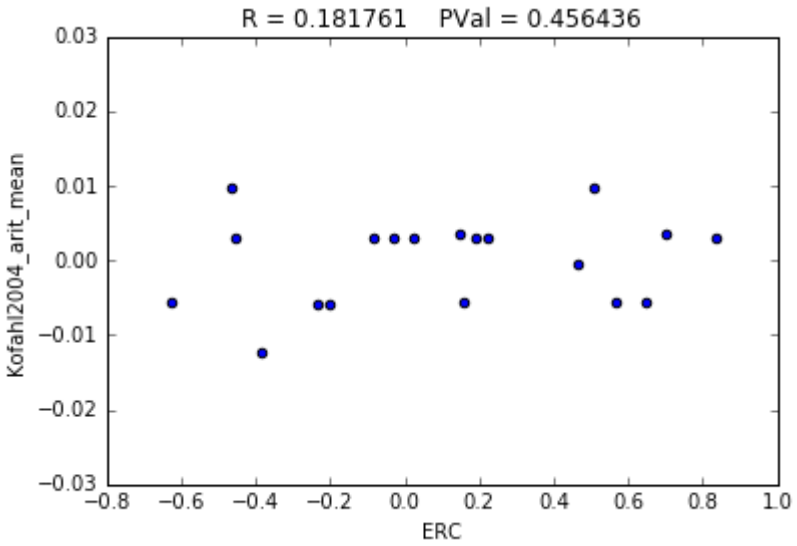
Empty scatter plots due to 0 or 1 (1 leaves only a diagonal entry) hessian parameter. Many protein pairs did not have overlapping interactions.

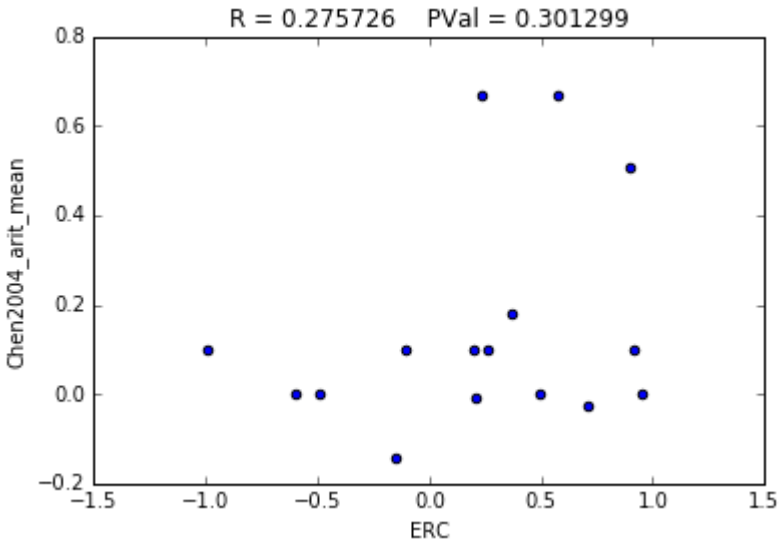





R = nan PVal = nan
Smallbone2011_arit_mean
ERC


R = nan PVal = nan
Ralser2007_arit_mean
ERC

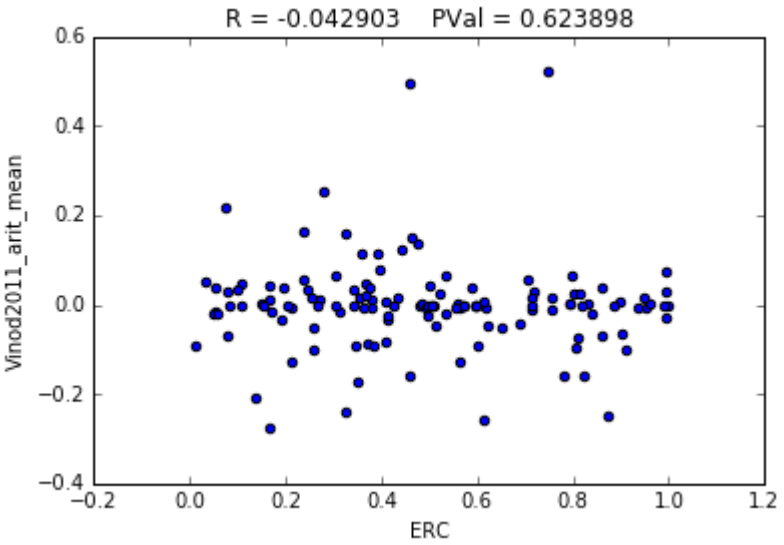

R = 0.181761 PVal = 0.456436
Kofahl2004_arit_mean
ERC

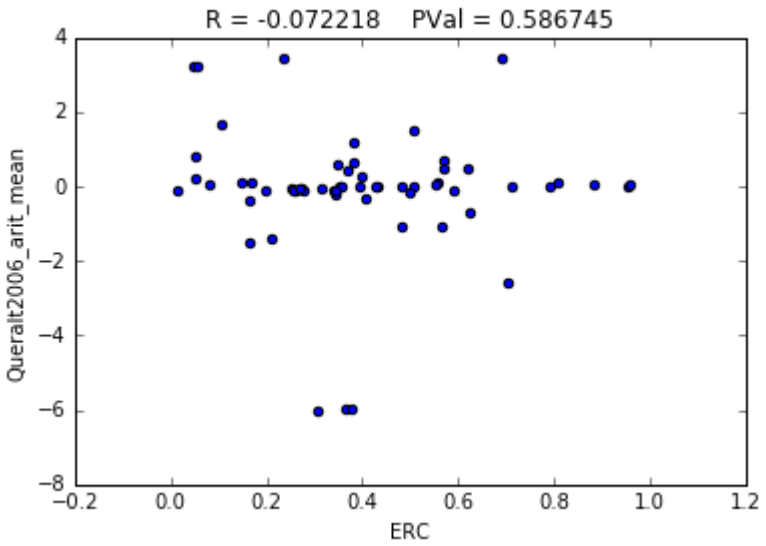

R = 0.275726 PVal = 0.301299
Chen2004_arit_mean
ERC
0.8
0.6
0.4
0.2
0.0
−0.2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

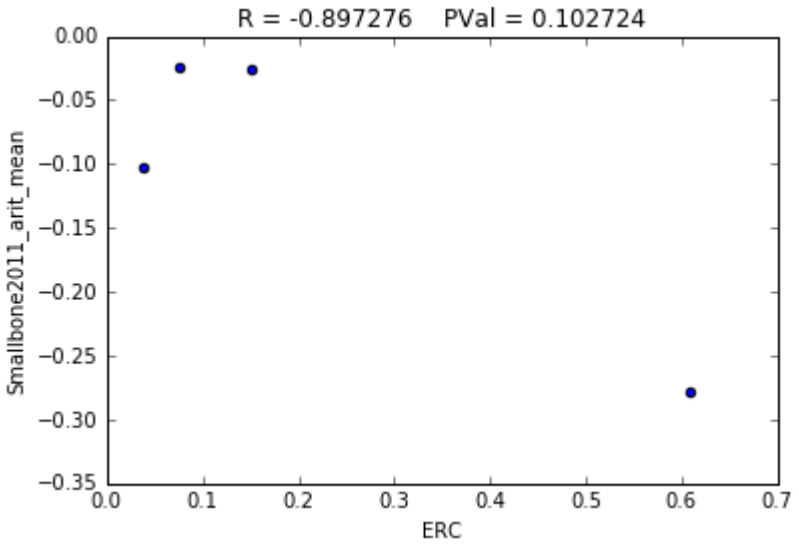
Union geometric absolute mean of hessian data vs erc

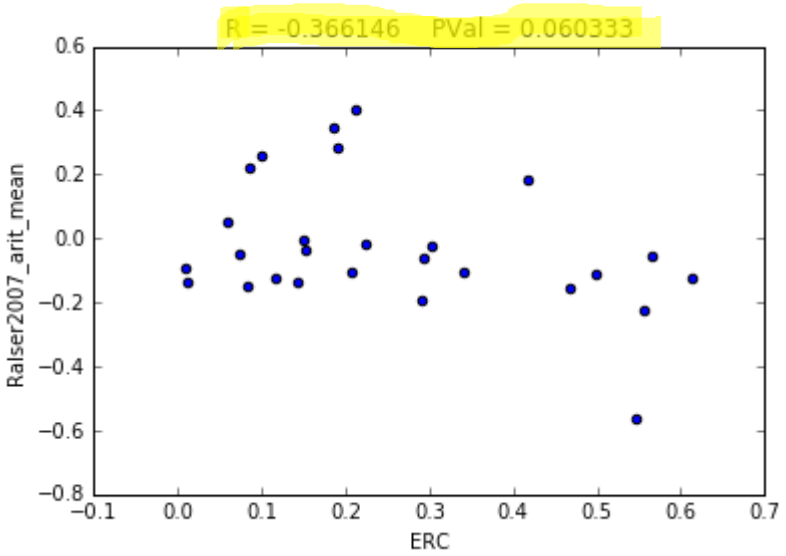
Monday, July 25, 2016 1:21 AM

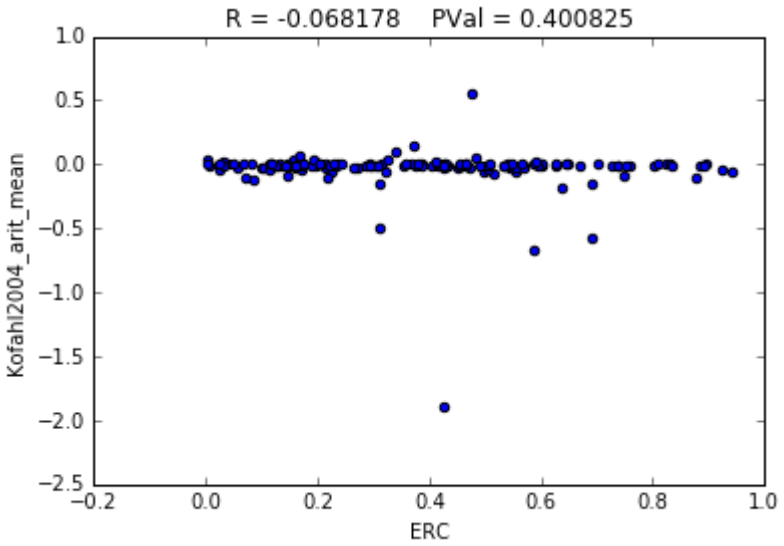
For all protein pairs, parameters contained by any reaction participated in by one of the pair, off diagonal hessian entries

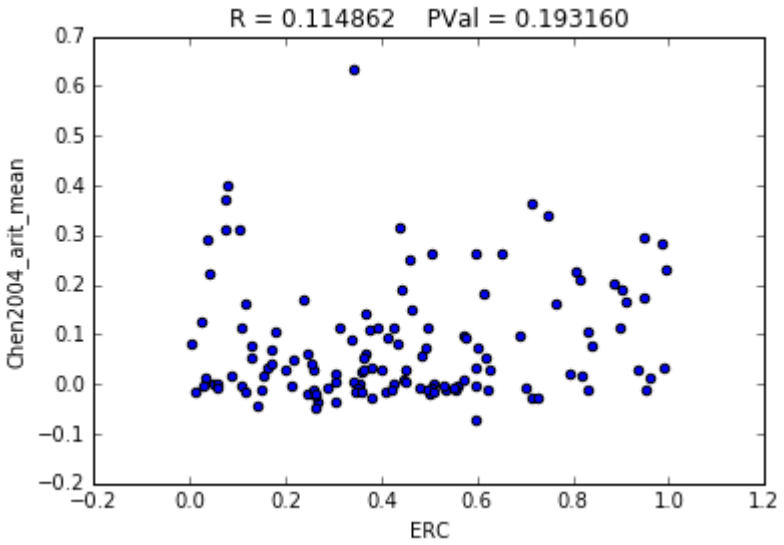





R = -0.897276 PVal = 0.102724
Smallbone2011_arit_mean
ERC


R = -0.366146 PVal = 0.060333
Ralser2007_arit_mean
ERC

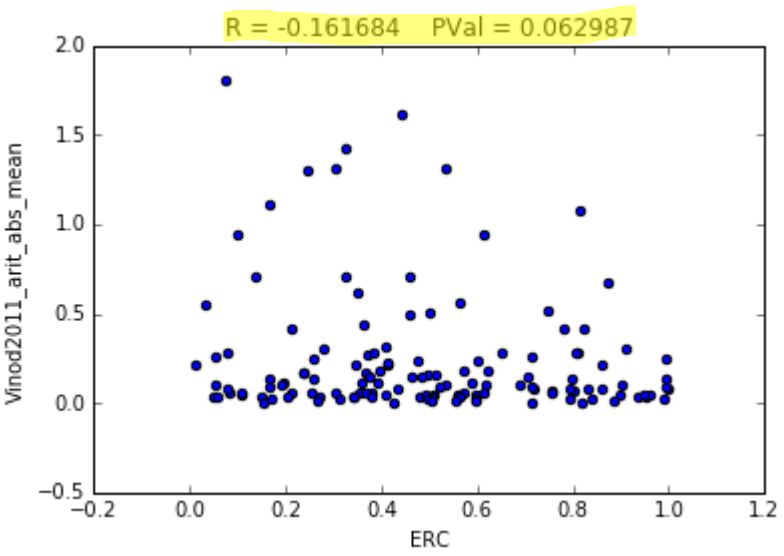

R = -0.068178 PVal = 0.400825
Kofahl2004_arit_mean
ERC

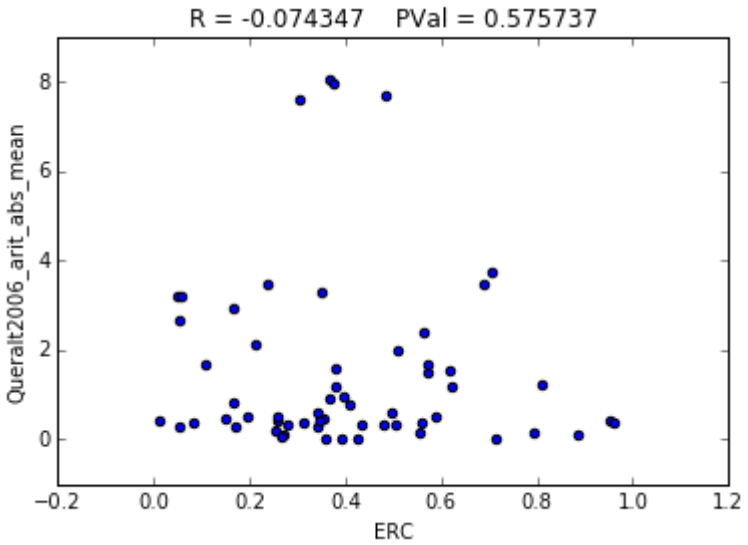

R = 0.114862 PVal = 0.193160
Chen2004_arit_mean
ERC
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2

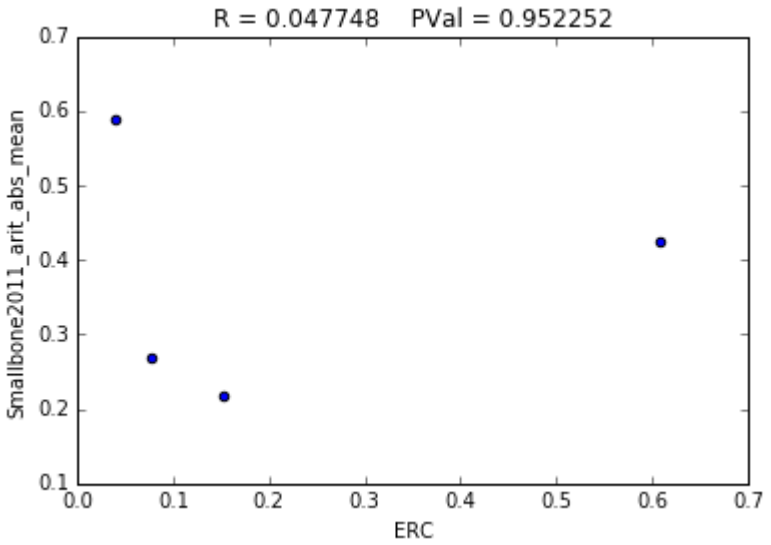
Union arithmetic absolute mean of hessian data vs erc

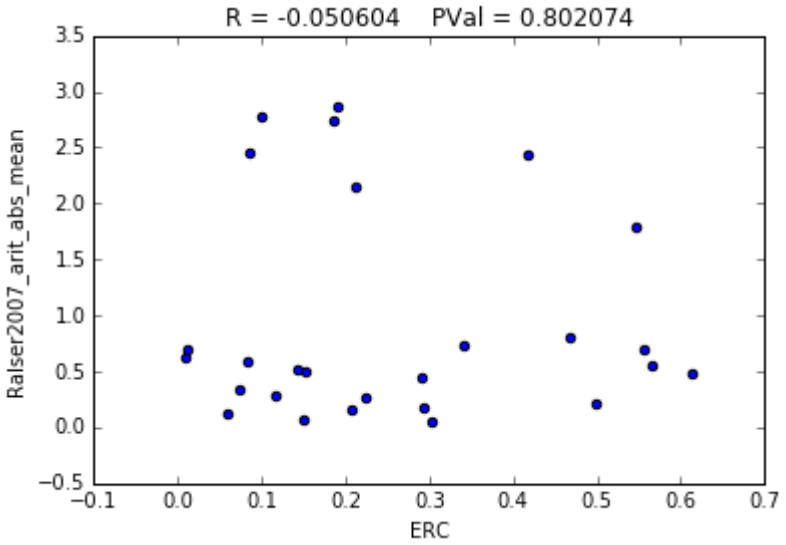
Monday, July 25, 2016 1:26 AM

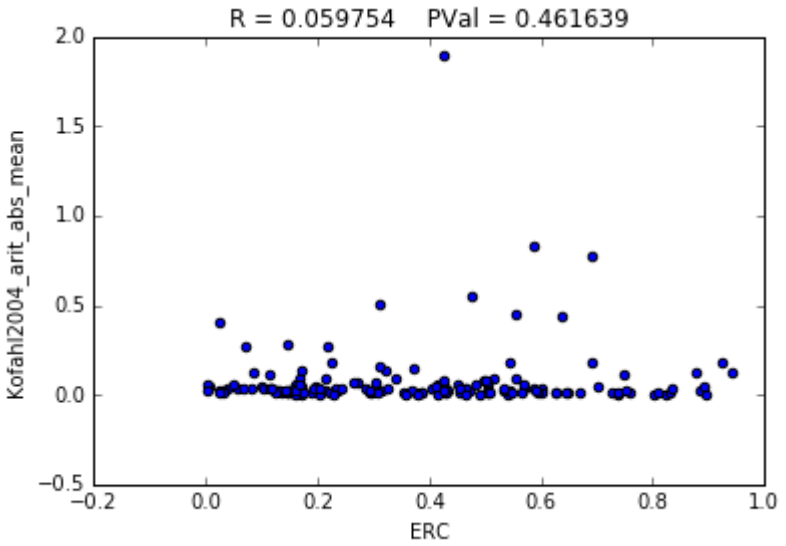
For all protein pairs, parameters contained by any reaction participated in by one of the pair, off diagonal hessian entries

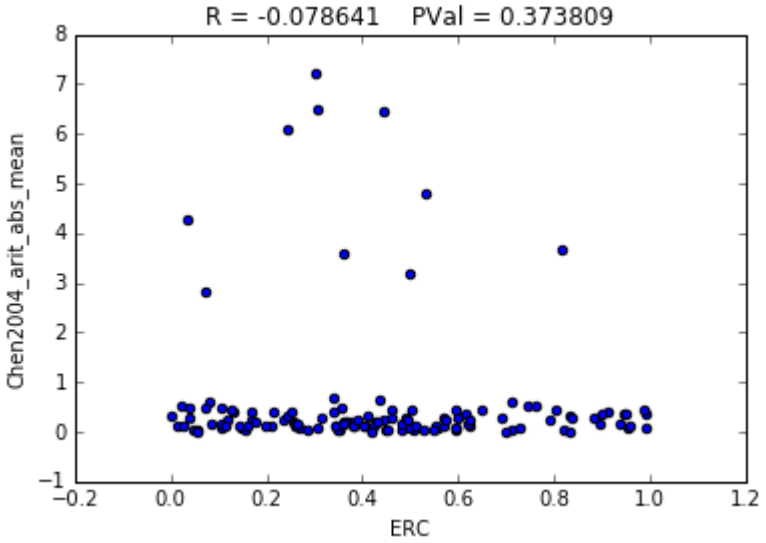





R = 0.047748 PVal = 0.952252
Smallbone2011_arit_abs_mean
ERC


R = -0.050604 PVal = 0.802074
Ralser2007_arit_abs_mean
ERC

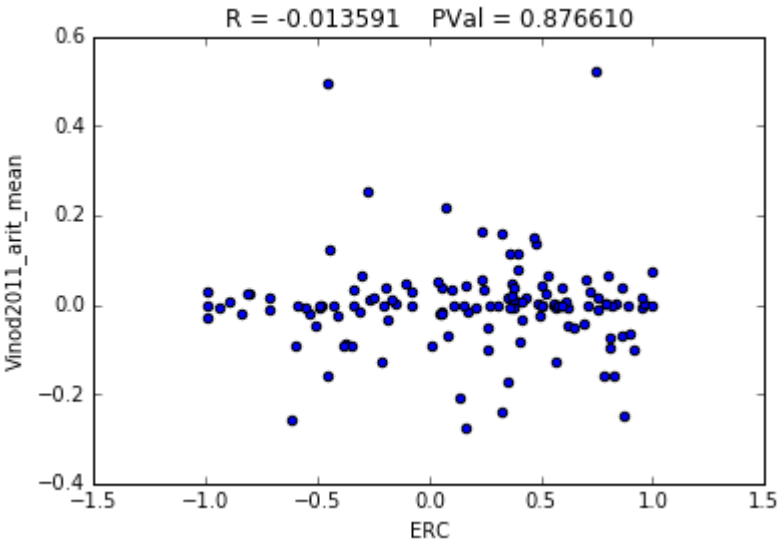

R = 0.059754 PVal = 0.461639
Kofahl2004_arit_abs_mean
ERC

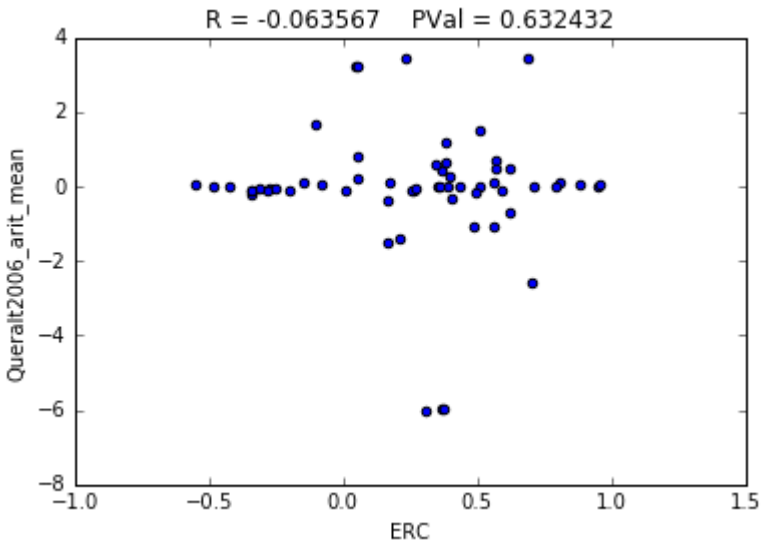

R = -0.078641 PVal = 0.373809
Chen2004_arit_abs_mean
ERC
8
7
6
5
4
3
2
1
0
−1
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2

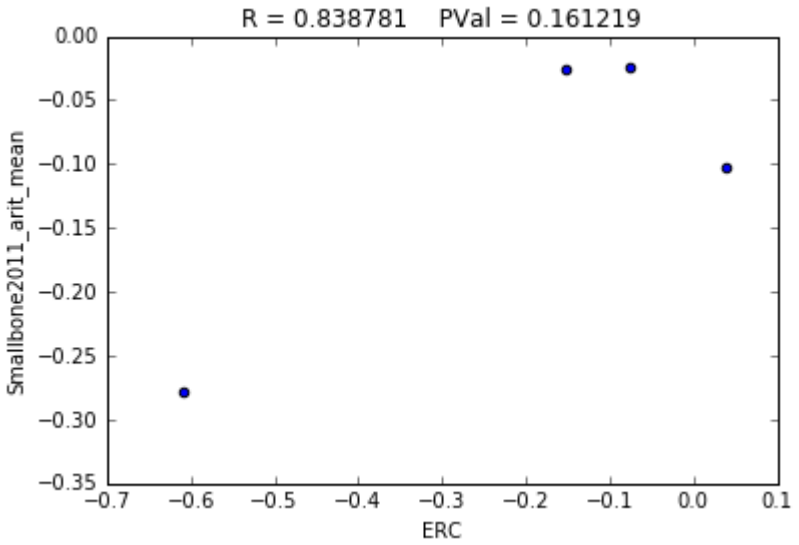
Union arithmetic mean of hessian data vs erc

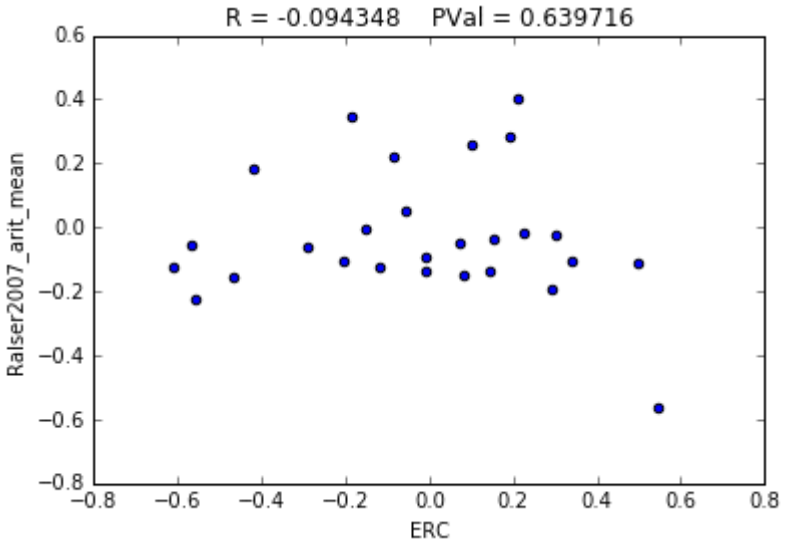
Monday, July 25, 2016 1:27 AM

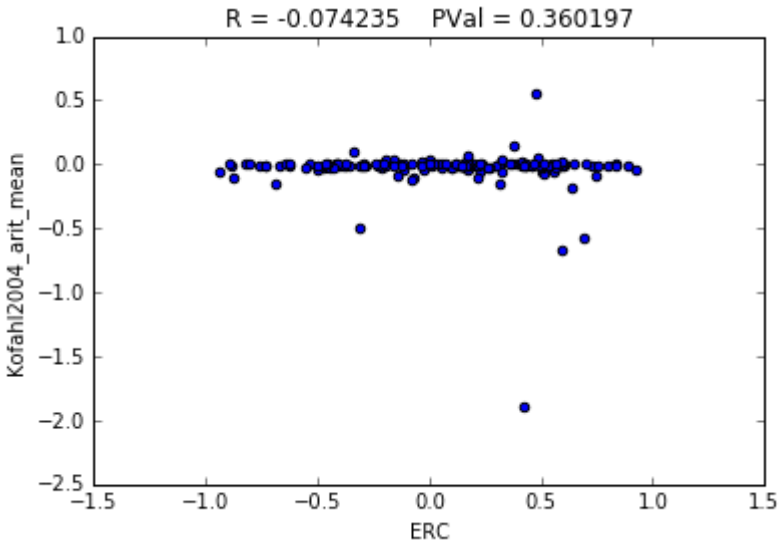
For all protein pairs, parameters contained by any reaction participated in by one of the pair, off diagonal hessian entries

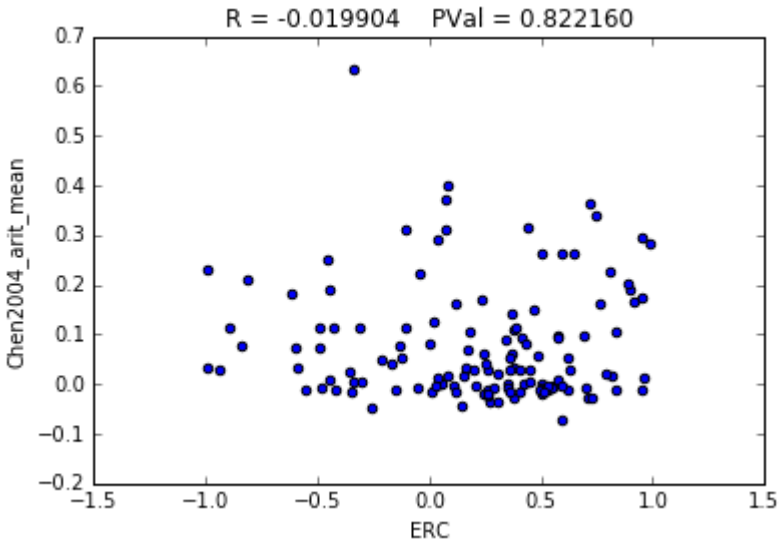





R = 0.838781 PVal = 0.161219
Smallbone2011_arit_mean
ERC


R = -0.094348 PVal = 0.639716
Ralser2007_arit_mean
ERC

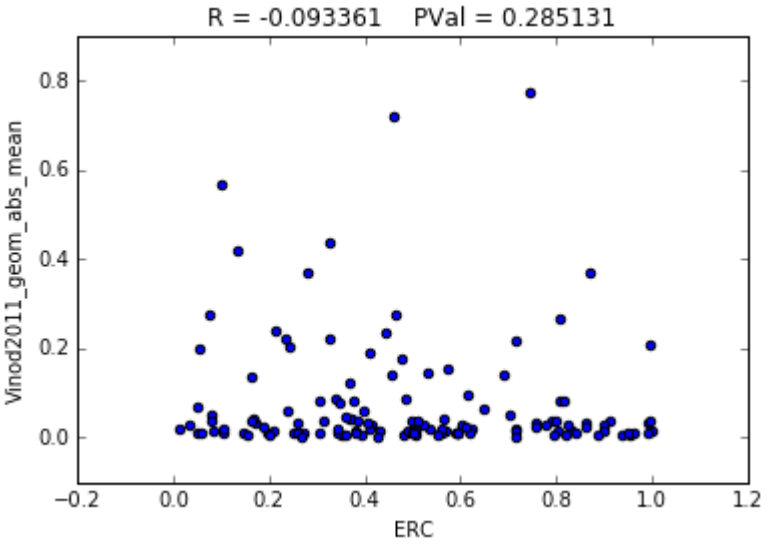

R = -0.074235 PVal = 0.360197
Kofahl2004_arit_mean
ERC

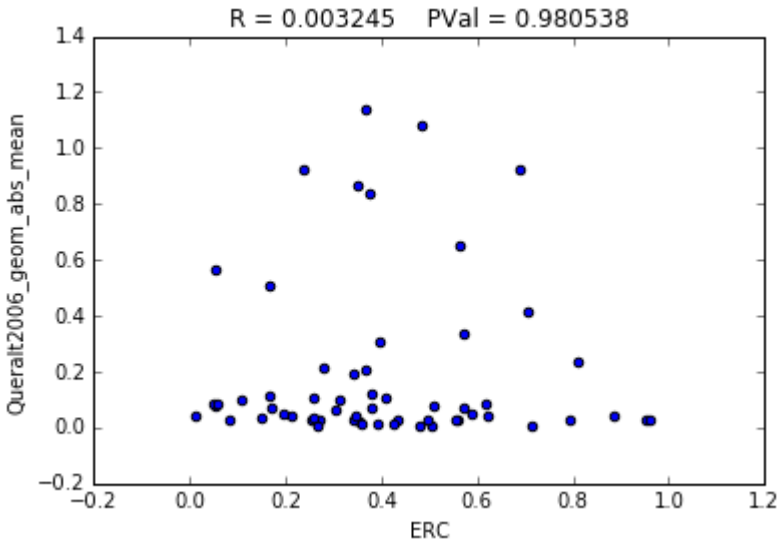

R = -0.019904 PVal = 0.822160
Chen2004_arit_mean
ERC
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

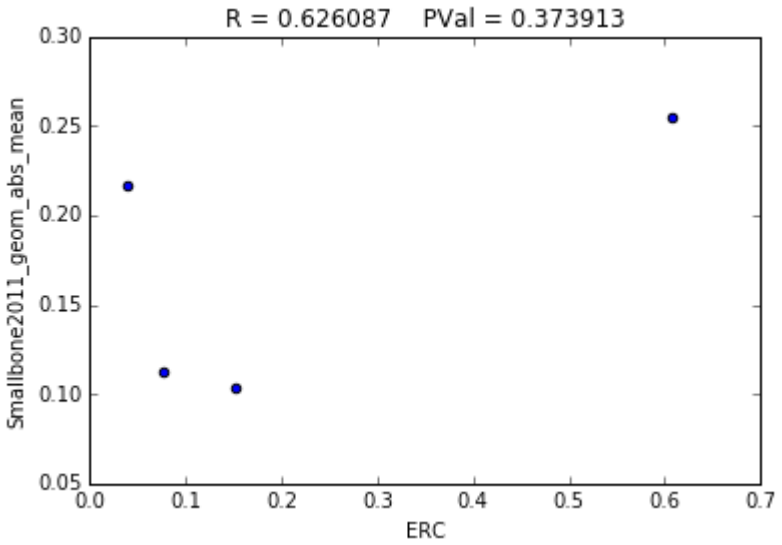
Union , w/ diagonals geometric abs mean

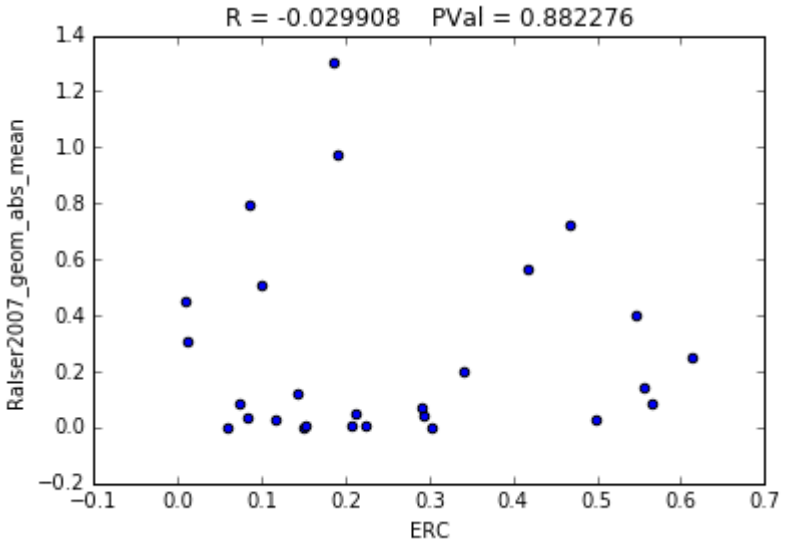
Monday, July 25, 2016 1:45 AM

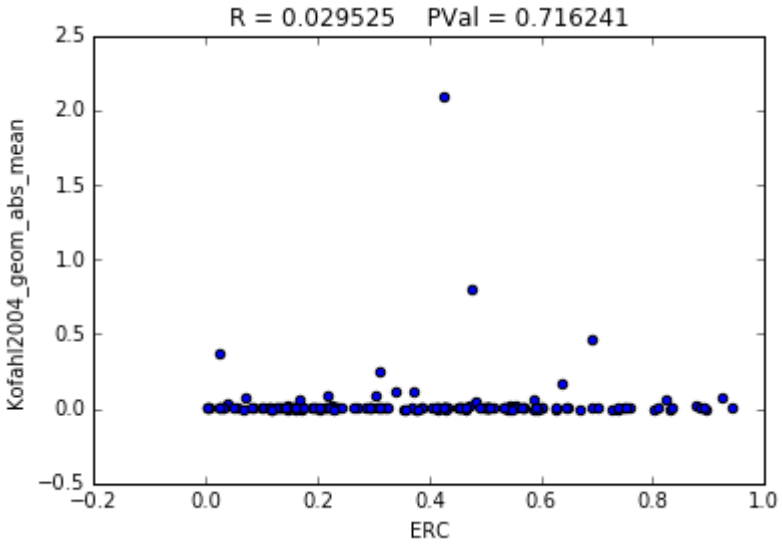
For all protein pairs, parameters contained by any reaction participated in by one of the pair, including diagonal hessian entries

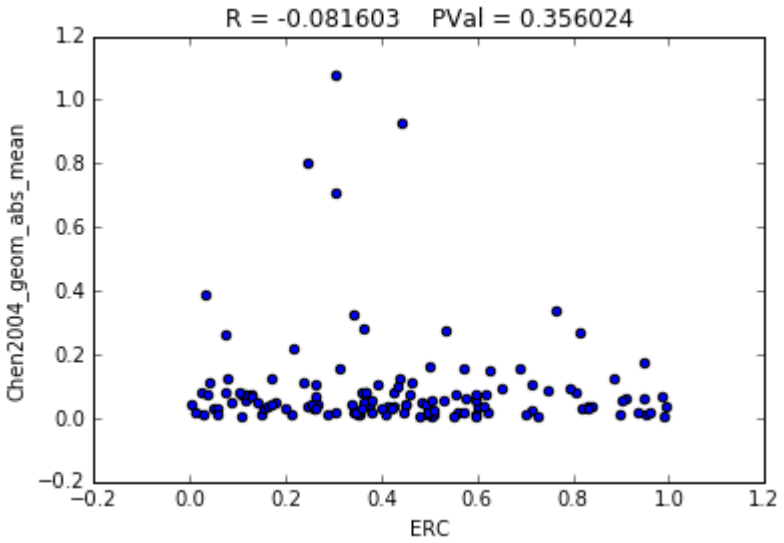





R = 0.626087 PVal = 0.373913
Smallbone2011_geom_abs_mean
ERC


R = -0.029908 PVal = 0.882276
Ralser2007_geom_abs_mean
ERC


R = 0.029525 PVal = 0.716241
Kofahl2004_geom_abs_mean
ERC


R = -0.081603 PVal = 0.356024
Chen2004_geom_abs_mean
ERC
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2

Interpretation

Monday, July 25, 2016 11:04 AM

P-values:

The low p-values seem to be just noticeably associated with high species counts in the phylogenetic trees for erc calculation

There are only ~7% below 5% P-value, increasing the number of species will probably yield more usable data

Distribution of hessians, means:

The distribution of hessians looks about normal, nothing unexpected. The distributions of the different means resemble one another

ERC distribution

There is a noticeable difference between the general ERC values and the subset of erc values between brian's interacting domains

The magnitude distribution shows this especially. This shows there is some signal, at least in the interaction data.

Shared reaction means:

Very little correlation. Any higher r values are influenced by outliers. P-values are very high. Some plots empty due
to either no shared reactions, or only one parameter shared (leaving only a diagonal, which isn't counted)

All three means show pretty similar patterns

Lots of flat lines near 0

Union reaction means:

More datapoints, similar to the shared reaction means

Union w/ diagonals

About the same as the others

Next

Is there another function we should be looking at to relate the influence to ERC?

Network properties

Meeting Notes:

Rank correlation
Log scale of correlation (expand values around 0)
ERC value signs

Look for papers/ ideas about relationships, datasets

Databases

Get vertebrate data
Diagram of system to understand (in papers cited)

Vertebrate Species

Friday, July 29, 2016 2:35 PM

Include Fish?

Brian:
~~Gallus gallus~~, Pan troglodytes, Danio rerio, Mus musculus, Rattus norvegicus, Bos taurus, Canis lupus familiaris, Homo sapiens

Nathan (https://files.acrobat.com/a/preview/f2b9388e-75bd-45f6-a021-855fe0fcd1ee)
Homo sapiens (human), Pongo pygmaeus abelii (orangutan), Macaca mulatta (rhesus macaque), Callithrix jacchus (marmoset), ~~Tarsius syrichta~~ (tarsier), Microcebus murinus (mouse lemur), Otolemur garnettii (bushbaby), Tupaia belangeri (tree shrew), Cavia porcellus (guinea pig), Dipodomys ordii (kangaroo rat), Mus musculus (mouse), Rattus norvegicus (rat), ~~Spermophilus tridecemlineatus~~ (squirrel), Oryctolagus cuniculus (rabbit), Ochotona princeps (pika), Vicugna pacos (alpaca), Sorex araneus (shrew), Bos taurus (cow), Tursiops truncatus (dolphin), Pteropus vampyrus (megabat), Myotis lucifugus (microbat), Erinaceus europaeus (hedgehog), Equus caballus (horse), Canis lupus familiaris (dog), Felis catus (cat), ~~Choloepus hoffmanni~~ (sloth), Echinops telfairi (tenrec), Loxodonta africana (elephant), ~~Procavia capensis~~ (rock hyrax), Dasypus novemcinctus (armadillo), Monodelphis domestica (opossum), ~~Macropus eugenii~~ (wallaby), and Ornithorhynchus anatinus (platypus)

Shared species	~~Not in refseq~~
Question	Found in refseq

30 of these species found in refseq

Name reference:

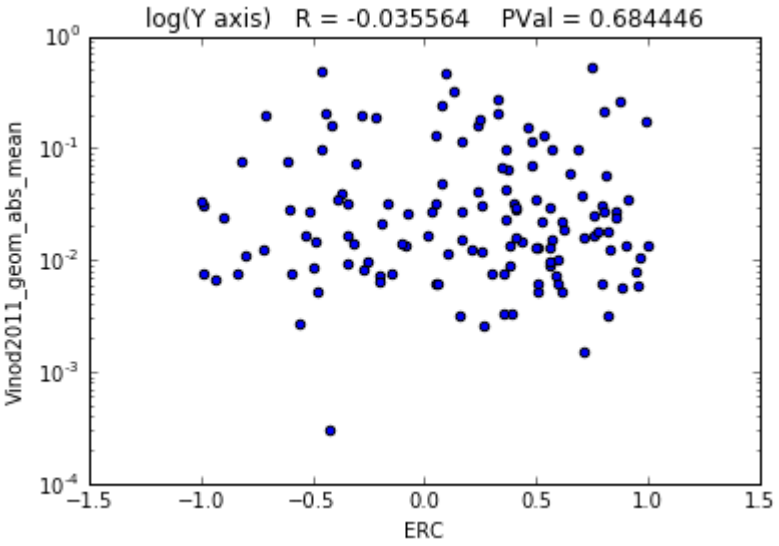
sapiens GCF_000001405.26_GRCh38_protein.faa.gz
troglodyt GCF_000001515.7_Pan_tro_3.0_protein.faa.gz
abelii GCF_000001545.4_P_pygmaeus_2.0.2_protein.faa.gz
musculus GCF_000001635.25_GRCm38.p5_protein.faa.gz
norvegicu GCF_000001895.5_Rnor_6.0_protein.faa.gz
africanaGCF_000001905.1_Loxafr3.0_protein.faa.gz
anatinus GCF_000002275.2_Ornithorhynchus_anatinus_5.0.1_protein.faa.gz
familiari GCF_000002285.3_CanFam3.1_protein.faa.gz
domestica GCF_000002295.2_MonDom5_protein.faa.gz
caballus GCF_000002305.2_EquCab2.0_protein.faa.gz
taurus GCF_000003055.6_Bos_taurus_UMD_3.1.1_protein.faa.gz
cuniculus GCF_000003625.3_OryCun2.0_protein.faa.gz
jacchus GCF_000004665.1_Callithrix_jacchus-3.2_protein.faa.gz
lucifugus GCF_000147115.1_Myoluc2.0_protein.faa.gz
porcellus GCF_000151735.1_Cavpor3.0_protein.faa.gz
vampyrus GCF_000151845.1_Pvam_2.0_protein.faa.gz
truncatus GCF_000151865.2_Ttru_1.4_protein.faa.gz
ordii GCF_000151885.1_Dord_2.0_protein.faa.gz
pacos GCF_000164845.2_Vicugna_pacos-2.0.2_protein.faa.gz
murinus GCF_000165445.1_Mmur_2.0_protein.faa.gz
araneusGCF_000181275.1_SorAra2.0_protein.faa.gz
garnettii GCF_000181295.1_OtoGar3_protein.faa.gz
catus GCF_000181335.2_Felis_catus_8.0_protein.faa.gz

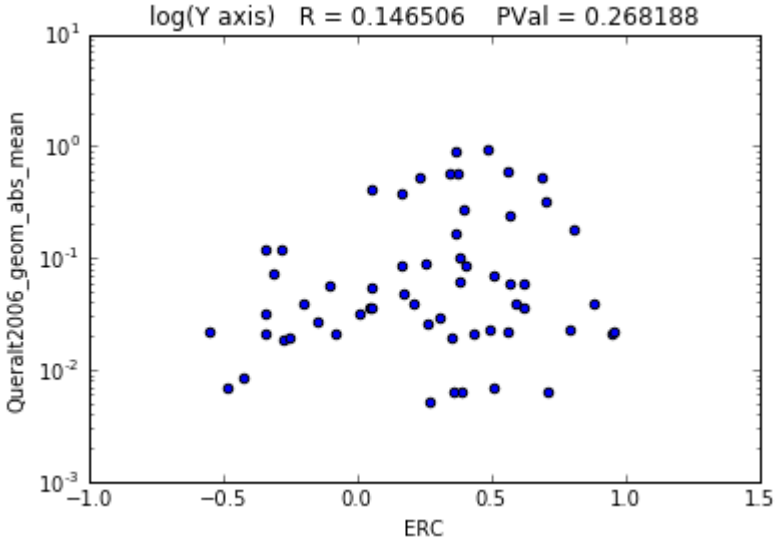
novemcinc GCF_000208655.1_Dasnov3.0_protein.faa.gz
princeps GCF_000292845.1_OchPri3.0_protein.faa.gz
europaeus GCF_000296755.1_EriEur2.0_protein.faa.gz
telfairi GCF_000313985.1_EchTel2.0_protein.faa.gz
belangeri GCF_000334495.1_TupChi_1.0_protein.faa.gz
mulatta GCF_000772875.2_Mmul_8.0.1_protein.faa.gz

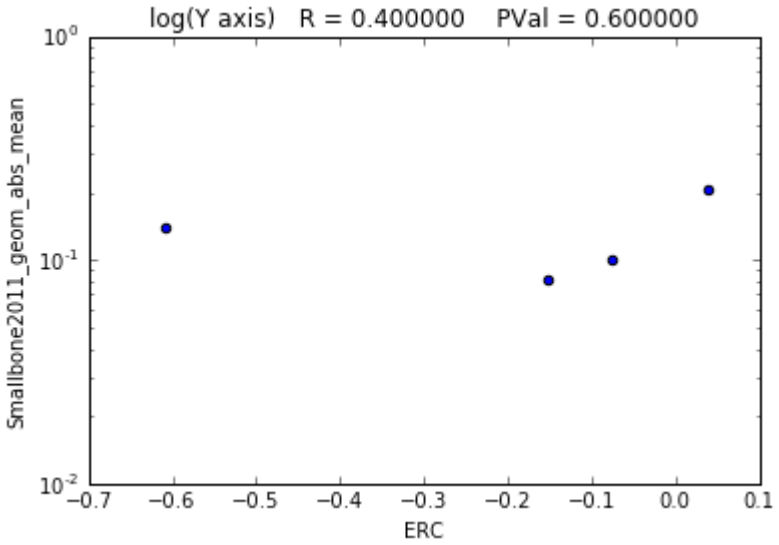
Union geometric absolute log mean of hessian data vs erc

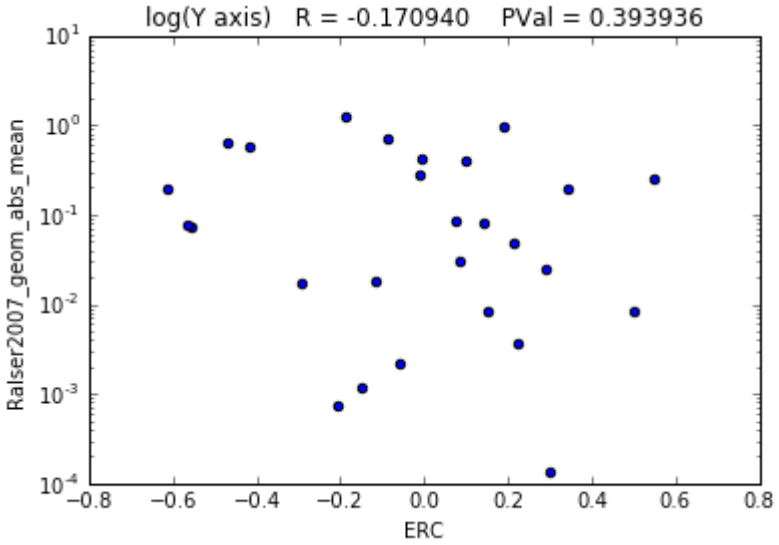
Monday, August 1, 2016 12:51 AM

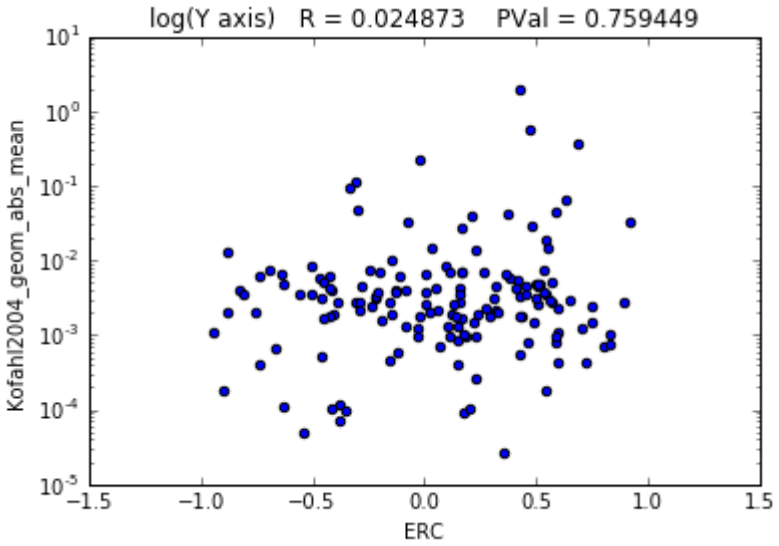
For all protein pairs, parameters contained by any reaction participated in by
one of the pair, off diagonal hessian entries. Mean log transformed. Spearman's correlation calculated

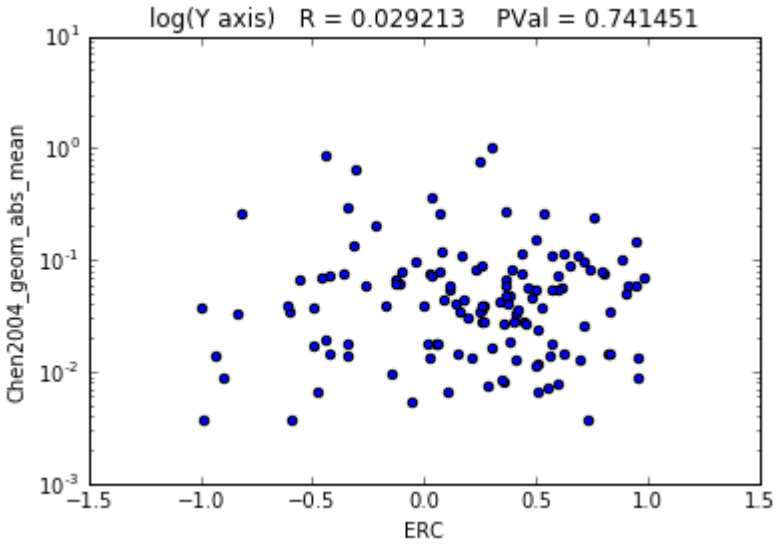





log(Y axis) R = 0.400000 PVal = 0.600000
Smallbone2011_geom_abs_mean
ERC


log(Y axis) R = -0.170940 PVal = 0.393936
Ralser2007_geom_abs_mean
ERC


log(Y axis) R = 0.024873 PVal = 0.759449
Kofahl2004_geom_abs_mean
ERC


log(Y axis) R = 0.029213 PVal = 0.741451
Chen2004_geom_abs_mean
ERC
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

Friday, July 29, 2016 4:19 PM

https://1drv.ms/b/s!AukW0kjQQyq1g-FAeDvg6GbjwPy4Ow

https://files.acrobat.com/a/preview/f2b9388e-75bd-45f6-a021-855fe0fcd1ee

Monday, August 1, 2016 12:27 AM

Things to look into:

Negative ERC values-- not really interprettable
Diagram of system
Nathan used codon bias in sequences as proxy for expression level. Are there other things that can be inferred from the sequences themselves?

$$\mathbf{H} = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1\, \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1\, \partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2\, \partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2\, \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n\, \partial x_1} & \dfrac{\partial^2 f}{\partial x_n\, \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

Interpretation

Monday, August 1, 2016 10:16 AM

Meeting notes:

Types of domains
Types of domain pairs
Specific examples of domains (Brian's)
Coinfluence function

Domain types and ERC

Monday, August 1, 2016 2:28 PM

	Domain type 1	Domain type 2	Domain type 3
Domain type 1	*		
Domain type 2	*	*	
Domain type 3	*	*	*

Domain type 1:
Gene1_2
Gene1_4
Gene27_2

Domain type 2:
Gene1_3
Gene4_2
Gene104_2

D1-D1	Gene1_2	Gene1_4	Gene27_2
Gene1_2	1		
Gene1_4	*	1	
Gene27_2	*	*	1

D1-D2	Gene1_2	Gene1_4	Gene27_2
Gene1_3	*		
Gene4_2	*	*	
Gene104_2	*	*	*

Domain type 1:
Gene1_2--Gene1_4: erc
Gene1_4--Gene27_2: erc
Gene27_2--Gene1_2: erc

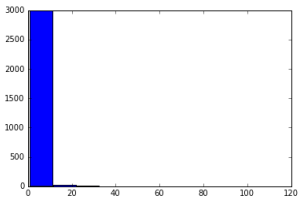
Each pair of domain types will have a collection of erc values like these.
The ones between the same domain type must exclude the diagonal.
Could have multiple descriptors, describing mean and standard deviation of values

Domain type--Gene--Domain# vs Gene_Domain#--Domain type
1st can easily make lists to feed into pandas. 2nd have to search for domain types, but can include in domain table(!)

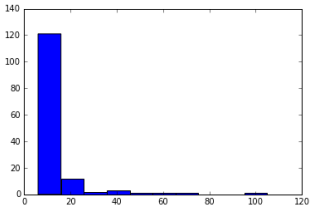
Domain type data

Monday, August 8, 2016 1:21 AM

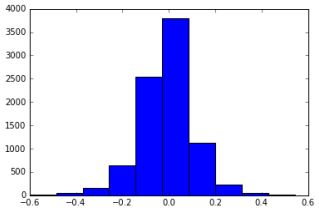
2888 unique domain types (with gene names) in yeast data
distribution of number of occurrences shown
mean: 2.0, median: 1.0, max: 105



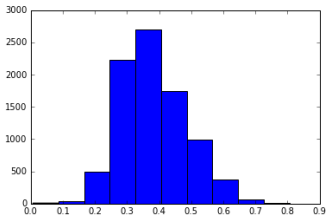
132 unique domain types with 5 or greater occurrences,
mean: 12.0, median: 8.0



Distribution of means between domain types
mean = -0.00819



Distribution of standard deviations between domain types
mean = 0.383



All hits above ERC value of 0.4:

('ABC2_membrane', 'Arf')	['erc mean: 0.49', 'erc std dev: 0.54']
('ABC2_membrane', 'FAD_binding_8')	['erc mean: 0.53', 'erc std dev: 0.40']
('Aldo_ket_red', 'adh_short')	['erc mean: 0.45', 'erc std dev: 0.35']
('Brix', 'Brix') Brix domain	['erc mean: 0.44', 'erc std dev: 0.15']
('Brix', 'zf-met')	['erc mean: 0.45', 'erc std dev: 0.32']
('Cation_ATPase_N', 'SET')	['erc mean: 0.41', 'erc std dev: 0.34']

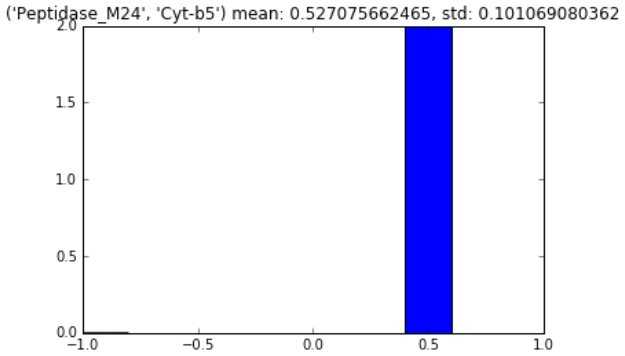
('Brix', 'Brix') Brix domain	['erc mean: 0.44', 'erc std dev: 0.15']
('Brix', 'zf-met')	['erc mean: 0.45', 'erc std dev: 0.32']
('Cation_ATPase_N', 'SET')	['erc mean: 0.41', 'erc std dev: 0.34']
('Cation_ATPase_N', 'adh_short')	['erc mean: 0.54', 'erc std dev: 0.39']
('FAD_binding_8', 'NAD_binding_6')	['erc mean: 0.53', 'erc std dev: 0.47']
('Ferric_reduct', 'FAD_binding_8')	['erc mean: 0.51', 'erc std dev: 0.56']
('Ferric_reduct', 'Glyco_transf_15')	['erc mean: 0.41', 'erc std dev: 0.46']
('Ferric_reduct', 'NAD_binding_6')	['erc mean: 0.54', 'erc std dev: 0.52']
('Glyco_transf_15', 'MutS_III')	['erc mean: 0.42', 'erc std dev: 0.44']
('Glyco_transf_15', 'NAD_binding_6')	['erc mean: 0.49', 'erc std dev: 0.35']
('HMG_box', 'FAD_binding_8')	['erc mean: 0.52', 'erc std dev: 0.50']
('HMG_box', 'NAD_binding_6')	['erc mean: 0.49', 'erc std dev: 0.56']
('HSP70', 'Snf7')	['erc mean: 0.41', 'erc std dev: 0.33']
Isocitrate/isopropylmalate dehydrogenase, Snf7	
('Iso_dh', 'Snf7')	['erc mean: 0.44', 'erc std dev: 0.28']
('MutS_III', 'MutS_III') MutS domain III	['erc mean: 0.49', 'erc std dev: 0.30']
('NAD_binding_6', 'NAD_binding_6')	['erc mean: 0.54', 'erc std dev: 0.47']
('PIR', 'Brix')	['erc mean: 0.40', 'erc std dev: 0.69']
('PIR', 'PIR')	['erc mean: 0.45', 'erc std dev: 0.33']
Yeast PIR protein repeat, Zinc-finger of C2H2 type	
('PIR', 'zf-met')	['erc mean: 0.44', 'erc std dev: 0.26']
('PUF', 'Snf7')	['erc mean: 0.46', 'erc std dev: 0.34']
Metallopeptidase family M24, Cytochrome b5-like Heme/Steroid binding domain	
('Peptidase_M24', 'Cyt-b5')	['erc mean: 0.53', 'erc std dev: 0.10']
('Peptidase_M24', 'PIR')	['erc mean: 0.45', 'erc std dev: 0.32']
('Snf7', 'Snf7')	['erc mean: 0.51', 'erc std dev: 0.17']
('Thioredoxin', 'Thioredoxin')	['erc mean: 0.49', 'erc std dev: 0.38']
short chain dehydrogenase	
('Thioredoxin', 'adh_short')	['erc mean: 0.45', 'erc std dev: 0.25']
('Transp_cyt_pur', 'Brix')	['erc mean: 0.42', 'erc std dev: 0.47']
('UBX', 'NAD_binding_6')	['erc mean: 0.41', 'erc std dev: 0.36']

Std dev <= 0.3

Means:	Standard deviations:
domain_type_means	domain_type_stds

Peptidase_M_24, Cyt-b5

Metallopeptidase family M24, Cytochrome b5-like Heme/Steroid binding domain



Peptidase M24, methionine aminopeptidase (IPR001714)

From <http://www.ebi.ac.uk/interpro/entry/IPR001714>

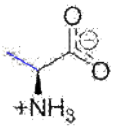
Methionine aminopeptidase (EC:3.4.11.18) (MAP) is responsible for the removal of the amino-terminal (initiator) methionine from nascent eukaryotic cytosolic and cytoplasmic prokaryotic proteins if the penultimate amino acid is small and uncharged. All MAP studied to date are monomeric proteins that require cobalt ions for activity.

Cytochrome b5-like heme/steroid binding domain(IPR001199)

From <http://www.ebi.ac.uk/interpro/entry/IPR001199>

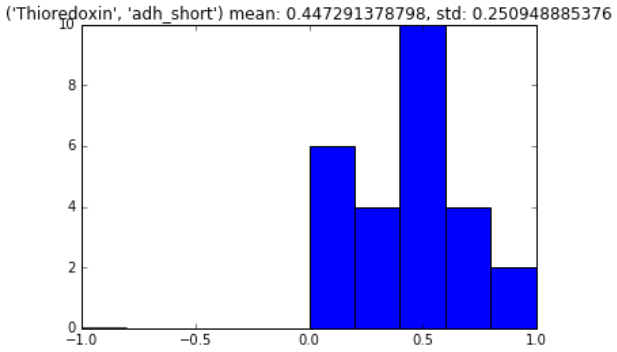
Cytochrome b5 is a membrane-bound hemoprotein which acts as an electron carrier for several membrane-bound oxygenases [PMID: 2752049]. There are two homologous forms of b5, one found in microsomes and one found in the outer membrane of mitochondria. Two conserved histidine residues serve as axial ligands for the heme group. The structure of a number of oxidoreductases consists of the juxtaposition of a heme-binding domain homologous to that of b5 and either a flavodehydrogenase or a molybdopterin domain.

>sp|P00167|2-134
AEQSDEAVKYYTLEEIQKHNHSKSTWLILHHKVYDLTKFLEEHPGGEEVLREQAGGDATE
NFEDVGHSTDAREMSKTFIIGELHPDDRPKLNKPPETLITTIDSSSSWWTNWVIPAISAV
AVALMYRLYMAED



Thioredoxin, adh_short

Thioredoxin, short chain dehydrogenase



Thioredoxin domain (IPR013766)

From <http://www.ebi.ac.uk/interpro/entry/IPR013766?q=Thioredoxin>

Thioredoxins [PMID: 3896121, PMID: 2668278, PMID: 7788289, PMID: 7788290] are small disulphide-containing redox proteins that have been found in all the kingdoms of living organisms. Thioredoxin serves as a general protein disulphide oxidoreductase. It interacts with a broad range of proteins by a redox mechanism based on reversible oxidation of two cysteine thiol groups to a disulphide, accompanied by the transfer of two electrons and two protons. The net result is the covalent interconversion of a disulphide and a dithiol. In the NADPH-dependent protein disulphide reduction, thioredoxin reductase (TR) catalyses the reduction of oxidised thioredoxin (trx) by NADPH using FAD and its redox-active disulphide; reduced thioredoxin then directly reduces the disulphide in the substrate protein

Short-chain dehydrogenase/reductase SDR(IPR002347)

From <https://www.ebi.ac.uk/interpro/entry/IPR002347>

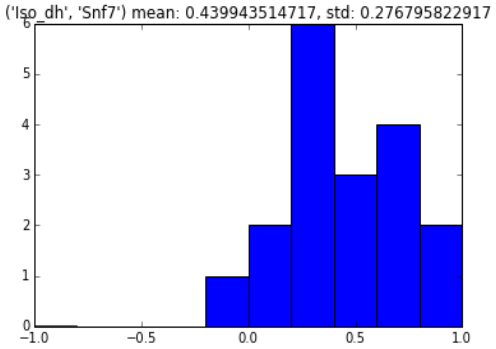
The short-chain dehydrogenases/reductases family (SDR) [PMID: 7742302, PMID: 25526675] is a very large family of enzymes, most of which are known to be NAD- or NADP-dependent oxidoreductases. As the first member of this family to be characterised was Drosophila alcohol dehydrogenase, this family used to be called [PMID: 2707261, PMID: 1889416,PMID: 1740120] 'insect-type', or 'short-chain' alcohol dehydrogenases. Most member of this family are proteins of about 250 to 300 amino acid residues. Most dehydrogenases possess at least 2 domains [PMID: 6789320], the first binding the coenzyme, often NAD, and the second binding the substrate. This latter domain determines the substrate specificity and contains amino acids involved in catalysis. Little sequence similarity has been found in the coenzyme binding domain although there is a large degree of structural similarity, and it has therefore been suggested that the structure of dehydrogenases has arisen through gene fusion of a common ancestral coenzyme nucleotide sequence with various substrate

specific domains

From <https://www.ebi.ac.uk/interpro/entry/IPR002347>

Iso_dh, Snf7

Isocitrate/isopropylmalate dehydrogenase, Snf7



Isopropylmalate dehydrogenase-like domain(IPR024084)

Short name: IsoPropMal-DH-like_dom

From <http://www.ebi.ac.uk/interpro/entry/IPR024084?q=iso_dh>

IDH is an important enzyme of carbohydrate metabolism which catalyses the oxidative decarboxylation of isocitrate into alpha-ketoglutarate [PMID: 2682654, PMID: 1939242]. IDH is either dependent on NAD+(EC:1.1.1.41) or on NADP+ (EC:1.1.1.42). In eukaryotes there are at least three isozymes of IDH: two are located in the mitochondrial matrix (one NAD+-dependent, the other NADP+-dependent), while the third one (also NADP+-dependent) is cytoplasmic. In Escherichia coli, the activity of a NADP+-dependent form of the enzyme is controlled by the phosphorylation of a serine residue; the phosphorylated form of IDH is completely inactivated.

IMDH (EC:1.1.1.85) catalyses the third step in the biosynthesis of leucine in bacteria and fungi, the oxidative decarboxylation of 3-isopropylmalate into 2-oxo-4-methylvalerate [PMID: 1748999, PMID: 7773180].

Snf7 family (IPR005024)

Short name: Snf7_fam

From <http://www.ebi.ac.uk/interpro/entry/IPR005024?q=snf7>

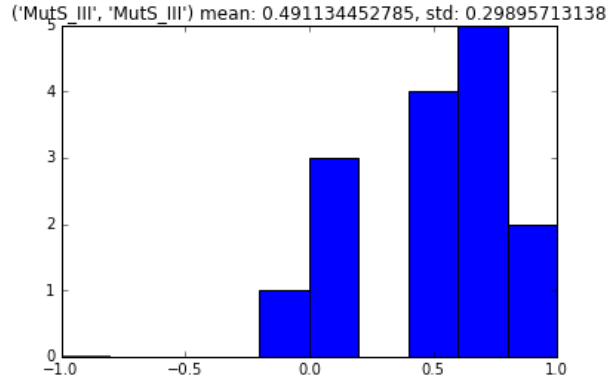
Snf7 family members are small coil-coiled proteins that share protein sequence similarity with budding yeast Snf7, which is part of the ESCRT-III

complex that is required for endosome-mediated trafficking via multivesicular body (MVB) formation and sorting [PMID: 15086794].

Proteins in this entry also includes human CHMPs (charged multivesicular body proteins), budding yeast Did4/Did2 and Arabidopsis vacuolar protein sorting-associated proteins.

MutS_III, MutS_III

MutS domain III, MutS domain III



DNA mismatch repair protein MutS, core (IPR007696)

Short name: DNA_mismatch_repair_MutS_core

Mismatch repair contributes to the overall fidelity of DNA replication and is essential for combating the adverse effects of damage to the genome. It involves the correction of mismatched base pairs that have been missed by the proofreading element of the DNA polymerase complex. The post-replicative Mismatch Repair System (MMRS) of Escherichia coli involves MutS (Mutator S), MutL and MutH proteins, and acts to correct point mutations or small insertion/deletion loops produced during DNA replication [PMID: 17919654]. MutS and MutL are involved in preventing recombination between partially homologous DNA sequences. The assembly of MMRS is initiated by MutS, which recognises and binds to mispaired nucleotides and allows further action of MutL and MutH to eliminate a portion of newly synthesized DNA strand containing the mispaired base [PMID: 17599803]. MutS can also collaborate with methyltransferases in the repair of O(6)-methylguanine damage, which would otherwise pair with thymine during replication to create an O(6)mG:T mismatch [PMID: 17951114]. MutS exists as a dimer, where the two monomers have different conformations and form a heterodimer at the structural level [PMID: 17426027]. Only one monomer recognises the mismatch specifically and has ADP bound. Non-specific major groove DNA-binding domains from both monomers embrace the DNA in a clamp-like structure. Mismatch binding induces ATP uptake and a conformational change in the MutS protein, resulting in a clamp that

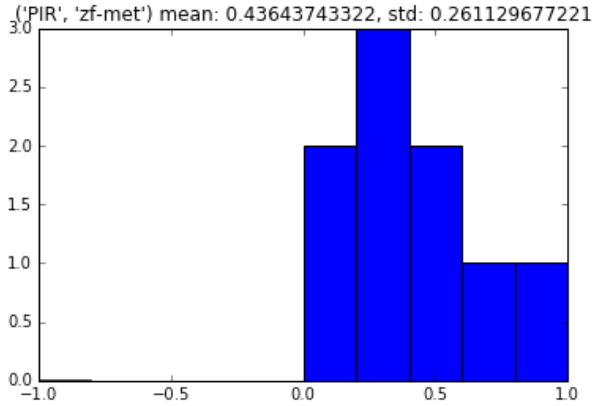
translocates on DNA.

From <http://www.ebi.ac.uk/interpro/entry/IPR007696?q=muts_III>

From <http://www.ebi.ac.uk/interpro/entry/IPR007696?q=muts_III>

PIR, zf-met

Yeast PIR protein repeat, Zinc-finger of C2H2 type



Yeast PIR protein repeat(IPR000420)

Short name: Yeast_PIR

A number of yeast cell wall glycoproteins are characterised by the presence of tandem repeats of a region of 18 to 19 residues [PMID: 8322511, PMID: 9301021].

From <http://www.ebi.ac.uk/interpro/entry/IPR000420?q=PIR>

Zinc-finger of C2H2 type

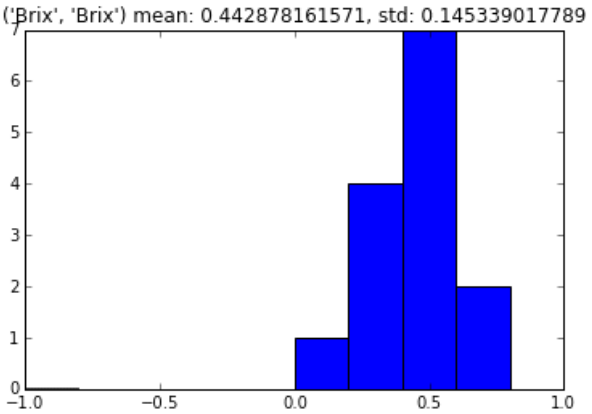
From <http://pfam.xfam.org/family/zf-met>

This is a zinc-finger domain with the CxxCx(12)Hx(6)H motif, found in multiple copies in a wide range of proteins from plants to metazoans. Some member proteins, particularly those from plants, are annotated as being RNA-binding.

From <http://pfam.xfam.org/family/zf-met>

Brix, Brix

Brix domain, Brix domain



Brix domain (IPR007109)

Short name: Brix

Analysis of the Brix (biogenesis of ribosomes in Xenopus) protein leaded to the identification of a region of 150-180 residues length, called the Brix domain, which is found in six protein families: one archaean family (I) including hypothetical proteins (one per genome); and five eukaryote families, each named according to a representative member and including close homologues of this prototype: (II) Peter Pan (D. melanogaster) and SSF1/2 (S.cerevisiae); (III) RPF1 (S. cerevisiae); (IV) IMP4 (S. cerevisiae); (V) Brix (X.laevis) and BRX1 (S. cerevisiae); and (VI) RPF2 (S.cerevisiae).

From <http://www.ebi.ac.uk/interpro/entry/IPR007109?q=brix>

Ribosome biogenesis protein BRX1 (IPR026532)

This family consists of BRX1 and homologues. In yeast, BRX1 is part of a complex that also includes RPF1, RPF2 and SSF1 or SSF2. It is required for biogenesis of the 60S ribosomal subunit [PMID: 11843177].

From <http://www.ebi.ac.uk/interpro/entry/IPR026532?q=brx1>

From <http://www.ebi.ac.uk/interpro/entry/IPR026532?q=brx1>

POU domain, class 6, transcription factor 2 (IPR033056)

Short name: POU6F2

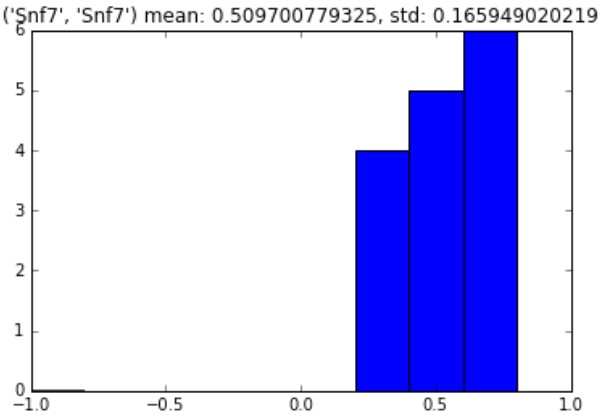
POU6F2, also known as retina-derived POU domain factor 1 (RPF1), is a transcription factor and a tumour suppressor associated with Wilms tumor (WT), a kidney malignancy of childhood characterised by highly heterogeneous genetic alterations [PMID: 17164647, PMID: 15459955]. In zebrafish it is involved retina regeneration [PMID: 25190811]. It has been shown to be involved in pituitary development [PMID: 24804940].

From <http://www.ebi.ac.uk/interpro/entry/IPR033056?q=rpf1>

From <http://www.ebi.ac.uk/interpro/entry/IPR033056?q=rpf1>

From <http://www.ebi.ac.uk/interpro/entry/IPR007109?q=brix>

Snf7, Snf7



Snf7 family (IPR005024)

Short name: Snf7_fam

From <http://www.ebi.ac.uk/interpro/entry/IPR005024?q=snf7>

Snf7 family members are small coil-coiled proteins that share protein sequence similarity with budding yeast Snf7, which is part of the ESCRT-III complex that is required for endosome-mediated trafficking via multivesicular body (MVB) formation and sorting [PMID: 15086794].

Proteins in this entry also includes human CHMPs (charged multivesicular body proteins), budding yeast Did4/Did2 and Arabidopsis vacuolar protein sorting-associated proteins.

Interpretation

Monday, August 8, 2016 1:04 PM

Heat map of subsequences of proteins in brians protiens/nathan's

Sliding window?

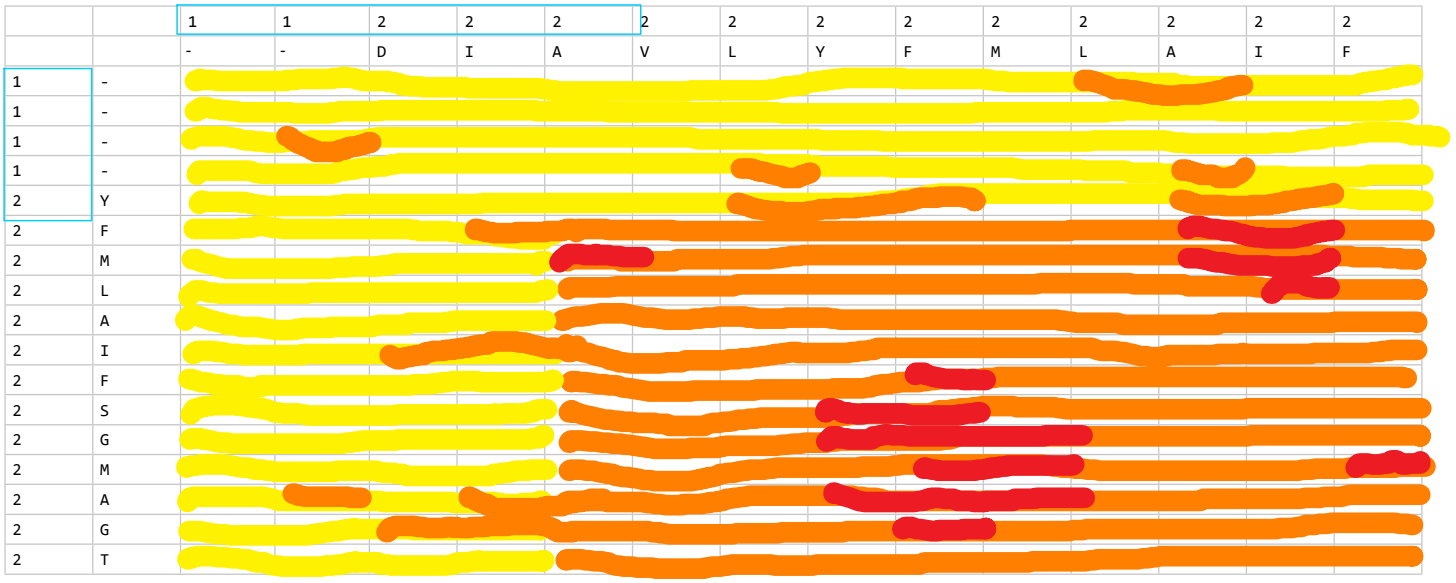
2 proteins, paper where interact (evolutionary signature) http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002452

Look at i2d interacting pairs for negative erc values (examples)

Possible windows

Tuesday, August 16, 2016 3:08 PM

Cerevisiae NP_009305.1--NP_009307.2 (altered for example)



		1	1	2	2	2	2	2	2	2	2	2	2	2	2
		-	-	D	I	A	V	L	Y	F	M	L	A	I	F
1	-														
1	-														
1	-														
1	-														
2	Y														
2	F														
2	M														
2	L														
2	A														
2	I														
2	F														
2	S														
2	G														
2	M														
2	A														
2	G														
2	T														

Sliding window of sequences or non-overlapping sections of sequence for comparison?

Sliding window gives more fine gradient between readings, but must be interpreted properly (each reading is results from the sequence on either side with the reading at the center, so this smears the value out, a matrix with one reading per amino acid gives the impression that that amino acid is responsible when it is hard to say that for sure. The readings at each end of the sequence would be cut short due to not being able to go any further)

Non-overlapping sections give a more accurate first impression, but make it hard to pinpoint the sections of the protein responsible for the reading.

Size of window: must be large enough to give accurate branch lengths in the phylogenetic tree, but small enough not to cut off too much of sequence ends.

Size of sequences(1000 unique yeast sequences with domain data):

Min = 49
1st quartile = 200
Median = 400
3rd quartile = 700
Max = 3000

Size of domains (manual sampling):

49, 94, 359, 59, 419, 1796, 111, 259, 134, 57, 94, 34, 6, 73, 107, 133, 27, 94, 20, 375, 163, 363, 660, 86, 98, 86, 100, 3, 83, 189, 62, 214, 13, 40, 146, 10, 87, 162, 373

min	3
q1	57
median	94
q3	189
max	1796

Assuming sequence length median 400 is close to mean, and a sliding window of 25 (so -50 for each sequence) 350 * 1,000 = 350,000 files if done this way, and 350,000 * 9 = 3,150,000 files if rechecking likelihood scores with already generated files. Compare this to 1,000 and 9,000 files for current method (will probably take forever unless I really spread it out, which is possible.

Running PAML last time (9,000 files) took ~6 hours. Say these smaller files take 25/400 =

0.0625; 0.0625 * 100 = 6.25%; 6 * 0.0625 = 0.375 hours. Assume there is a minimum to process each file by rounding up to 0.5 hours. 0.5 * 350,000/9,000 = 19.4444 hours; and 0.5 * 3,150,000/9,000 = 175 hours. 175 / 24 = 7.2917 days.

Assuming median 400 is close to mean, and a non-overlapping window of 25. 400 / 25 * 1,000 = 16,000 files, and 16,000 * 9 = 144,000 files. Using above calculations 0.5 * 144,000/9,000 = 8 hours.

25 feels like a good minimum for finding differences between similar sequences, still may not be enough. Compared to 1st quartile domain size, this would give at least 2 readings per domain on most sequences, probably acceptable for now.

Some readings will be between domains, but at least one will be completely on the domain in most cases)
Going to have to take a census [rounded mean] to see what domain to call it …
Although, might need even more fine grained to find anything interesting…
What size would be worthwhile? Maybe 15. 15 feels too small for PAML though. Maybe 20. It depends on what we're trying to get out of it. The location of interaction?
We already have everything divided up by domains. So the only point in doing a heat map would be to get a more detailed reading. Therefore the majority of domains in question must be subdivided at least into two, but three would be better, which gives subdivisions of ~20 for 75% of domains being divided into 3.

Another option would be to do the sliding windows, but only for a subset of the data, the subset in Brian's data. This may be ideal, at least if there is structural data on these, which there probably is… I need to check.

I'm finding some stuff, there should be enough to work with. It's easy to partition out Brian's stuff too. Should do it with Brian's annotations and pfam ones

Negative ERC

Tuesday, August 16, 2016 6:34 PM

100742	BioGrid_Yeast	SIN3	CHA1	-0.56631
100743	DIP_Yeast	PPH21	CHA1	-0.37371
100751	IntAct_Yeast	SSE1	CHA1	-0.37385
57419	DIP_Yeast	CHA1	CMP2	-0.20469

Standard Name

CHA1 [1]

Systematic Name

YCL064C

SGD ID

S000000569

Feature Type

ORF , Verified

Description

Catabolic L-serine (L-threonine) deaminase; catalyzes the degradation of both L-serine and L-threonine; required to use serine or threonine as the sole nitrogen source, transcriptionally induced by serine and threonine [1] [2]

Name Description

Catabolism of Hydroxy Amino acids [1]

From <http://www.yeastgenome.org/locus/S000000569/overview>

Standard Name

SIN3 [1]

Systematic Name

YOL004W

SGD ID

S000005364

Aliases

CPE1 [2] , GAM2 [3] , RPD1 [4] [5] , SDI1 [4] [6] , SDS16 [7] [8] , UME4 [4] [9]

Feature Type

ORF , Verified

Description

Component of both the Rpd3S and Rpd3L histone deacetylase complexes; involved in transcriptional repression and activation of diverse processes, including mating-type switching and meiosis; involved in the maintenance of chromosomal integrity [10] [11] [12]

Name Description

Switch INdependent

Dts	Molecule 'A'	Links 'A'	Molecule 'B'	Links 'B'	Interaction Detection Method	Interaction AC	Source Database	
		CHA1	P25379 EBI-3804607	SSE1	P32589 EBI-8648	tandem affinity purification	EBI-3804612	IntA

From <http://www.ebi.ac.uk/intact/interactions?conversationContext=8>

Interaction between molecules that may participate in formation of one, but possibly more, physical complexes. Often describes a set of molecules that are co-purified in a single pull-down or coimmunoprecipitation but might participate in formation of distinct physical complexes sharing a common bait.

From <http://www.ebi.ac.uk/intact/interaction/EBI-3804612?conversationContext=9&kmr=false>

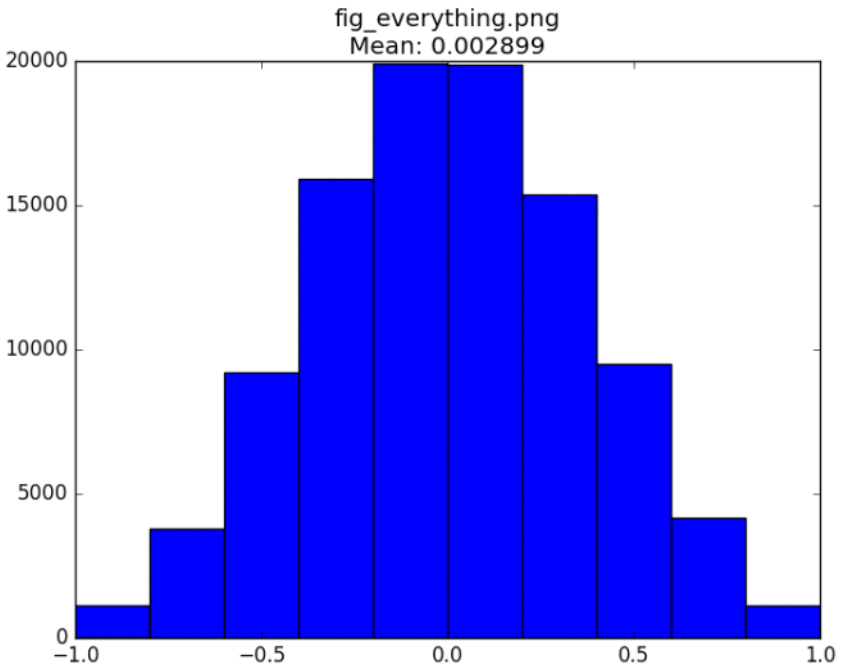
Negative ERC 2

Wednesday, August 17, 2016 9:49 PM

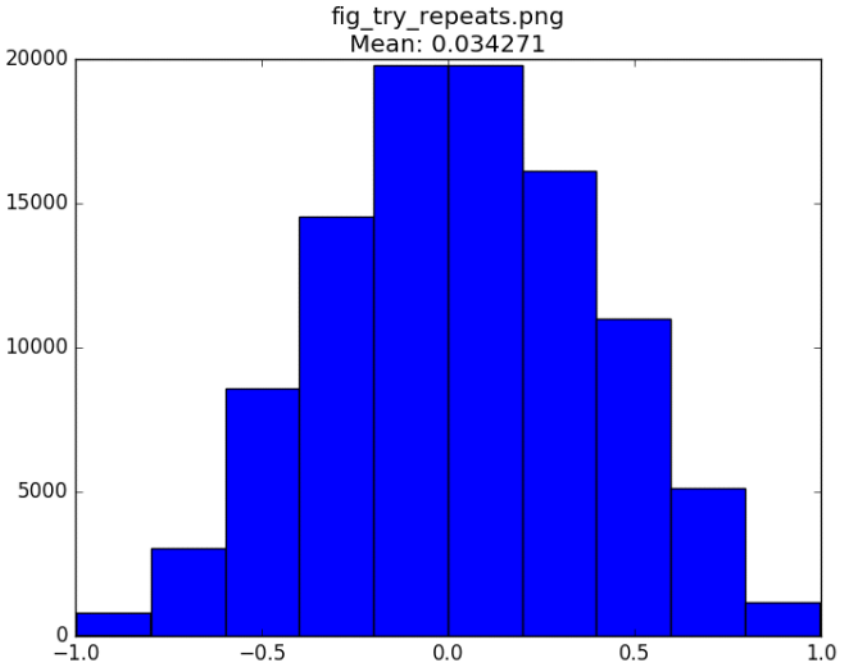
Can't replicate results found here with script: interact_analyze

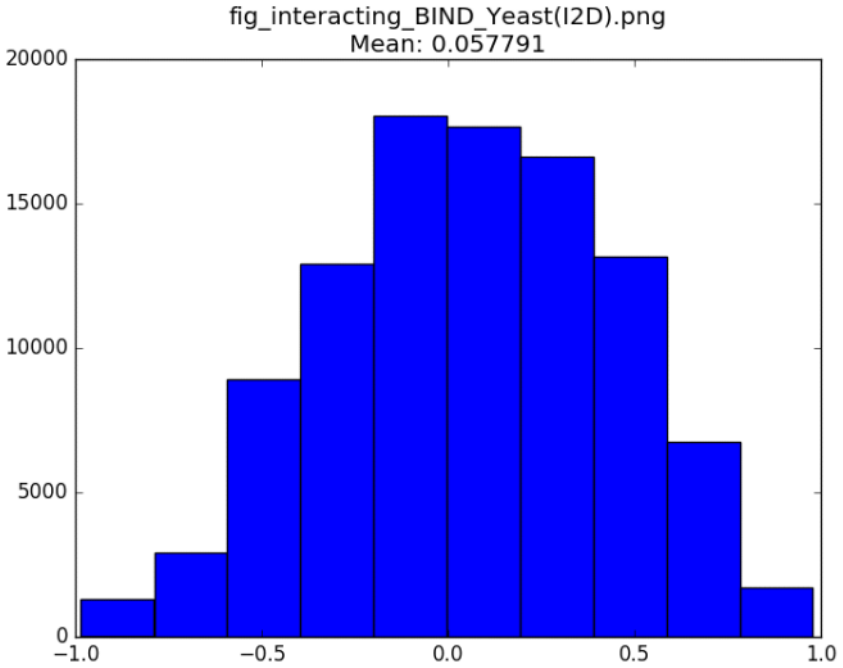
Here's what I'm now getting:

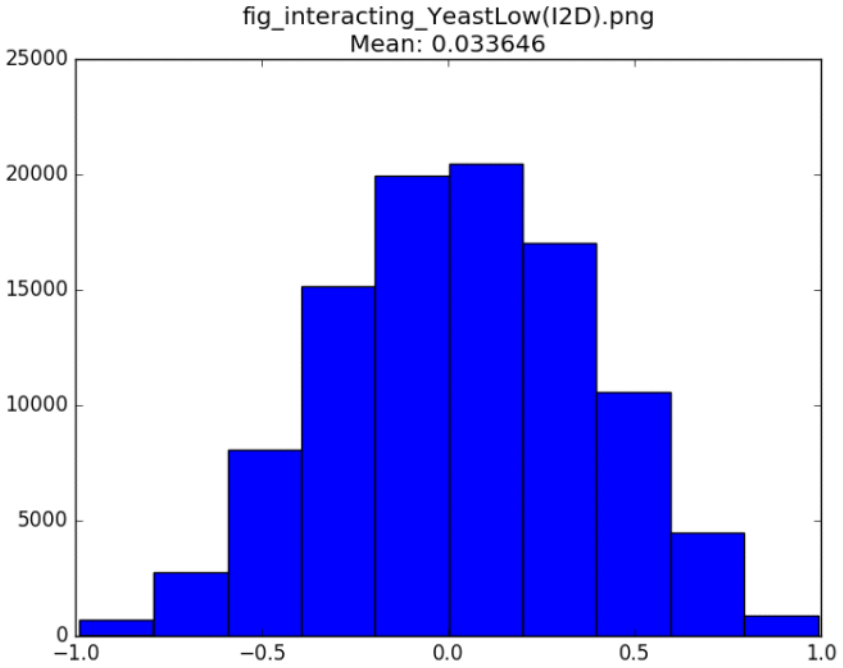
All ERC vals without filtering out those not interacting according to I2D:

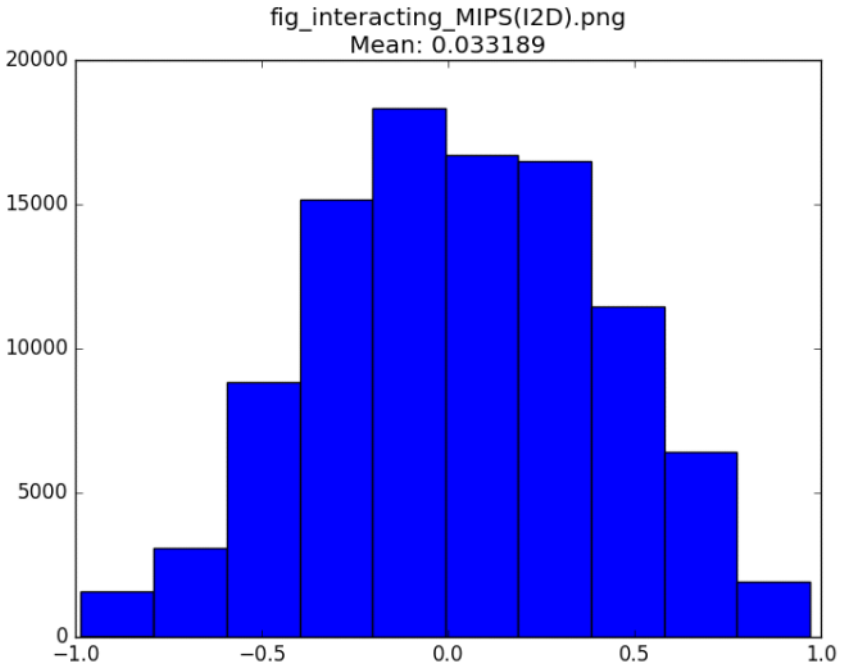


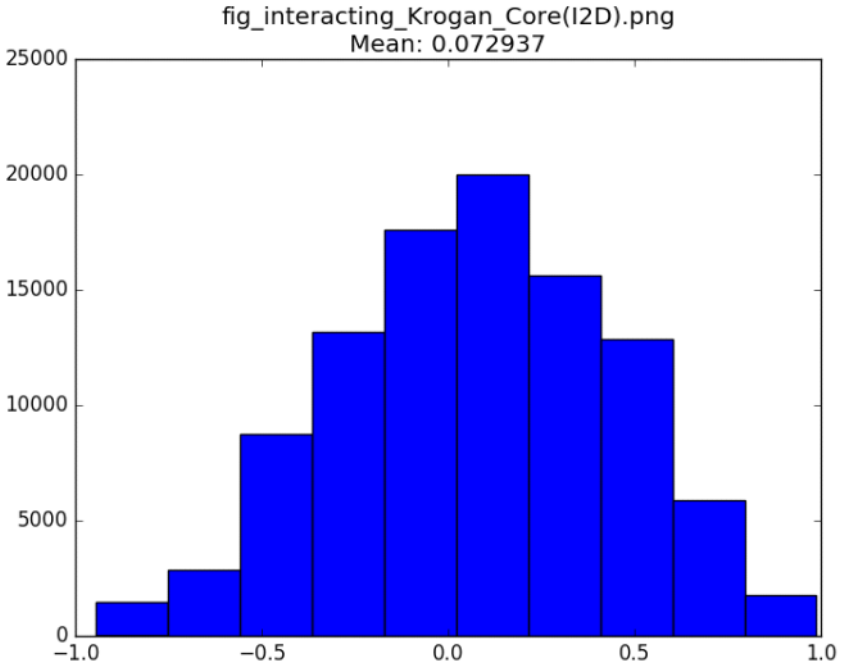
ERC vals filtering by interaction according to I2D (1st all sources [with repeats], then each source:

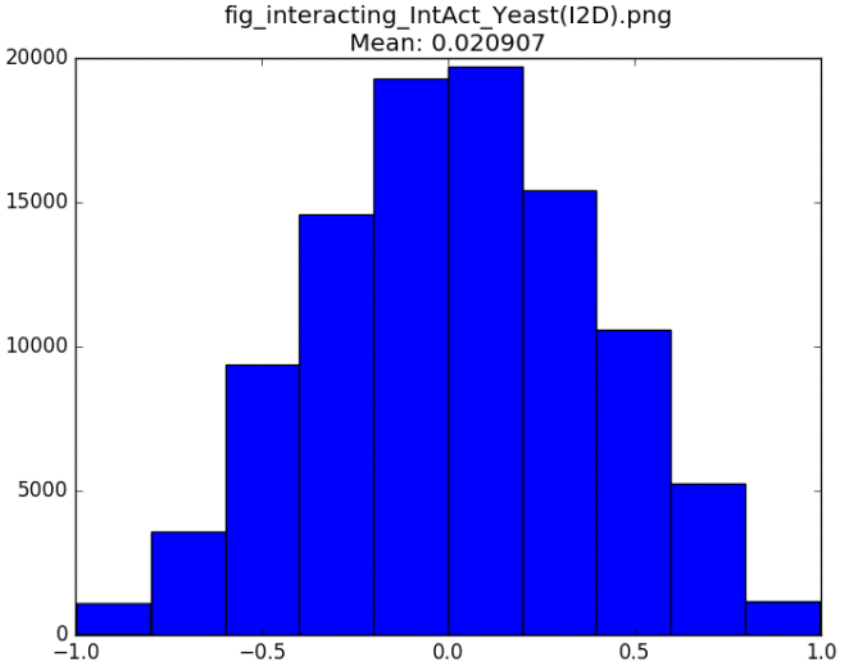

fig_try_repeats.png
Mean: 0.034271
20000
15000
10000
5000
0
−1.0
−0.5
0.0
0.5
1.0

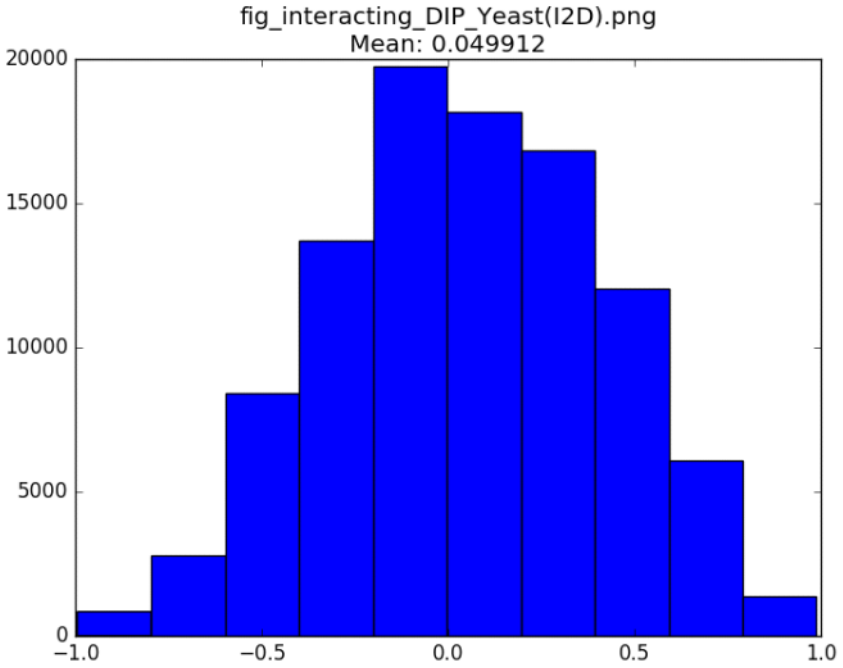

fig_interacting_BIND_Yeast(I2D).png
Mean: 0.057791
20000
15000
10000
5000
0
−1.0
−0.5
0.0
0.5
1.0

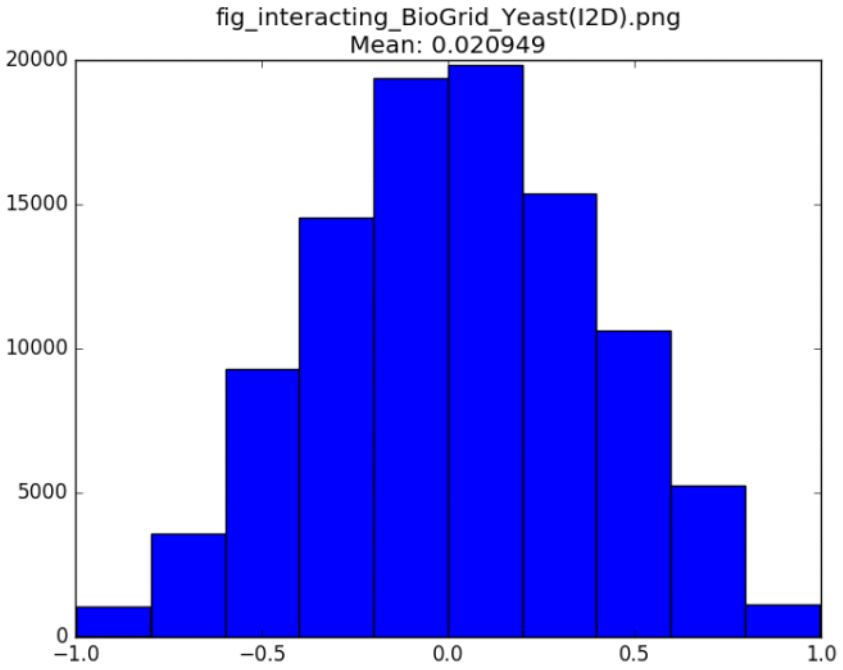

fig_interacting_YeastLow(I2D).png
Mean: 0.033646
25000
20000
15000
10000
5000
0
−1.0
−0.5
0.0
0.5
1.0


fig_interacting_MIPS(I2D).png
Mean: 0.033189
20000
15000
10000
5000
0
−1.0
−0.5
0.0
0.5
1.0


fig_interacting_Krogan_Core(I2D).png
Mean: 0.072937
25000
20000
15000
10000
5000
0
−1.0
−0.5
0.0
0.5
1.0


fig_interacting_IntAct_Yeast(I2D).png
Mean: 0.020907
20000
15000
10000
5000
0
−1.0
−0.5
0.0
0.5
1.0





Using only data with all 13 species:

Everything (no filtering):

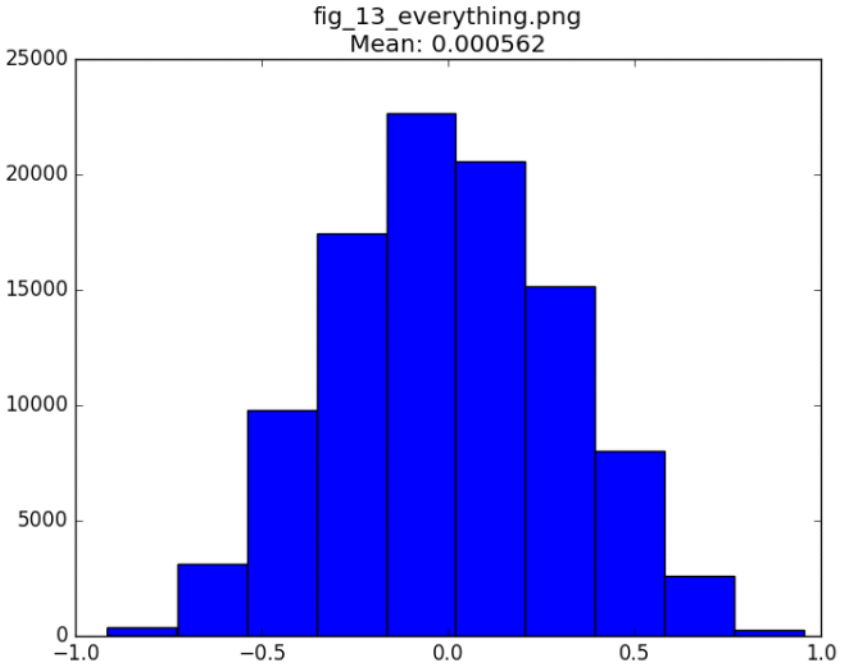


fig_13_everything.png
Mean: 0.000562
25000
20000
15000
10000
5000
0
−1.0
−0.5
0.0
0.5
1.0

Filtering:

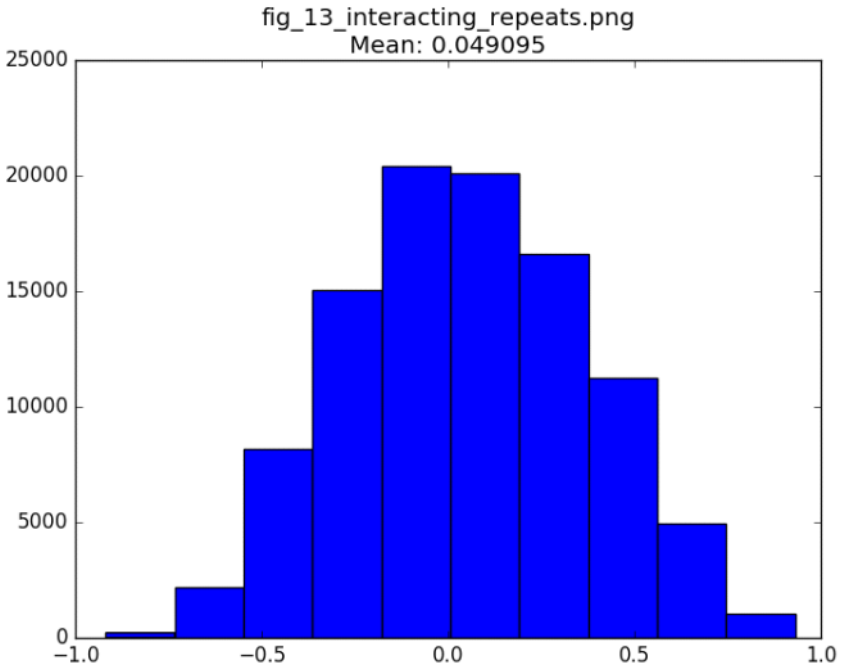


fig_13_interacting_repeats.png
Mean: 0.049095
25000
20000
15000
10000
5000
0
−1.0
−0.5
0.0
0.5
1.0

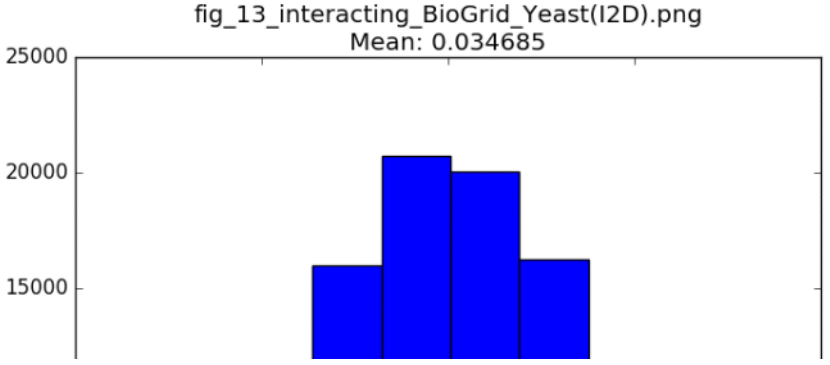


fig_13_interacting_BioGrid_Yeast(I2D).png
Mean: 0.034685
25000
20000
15000

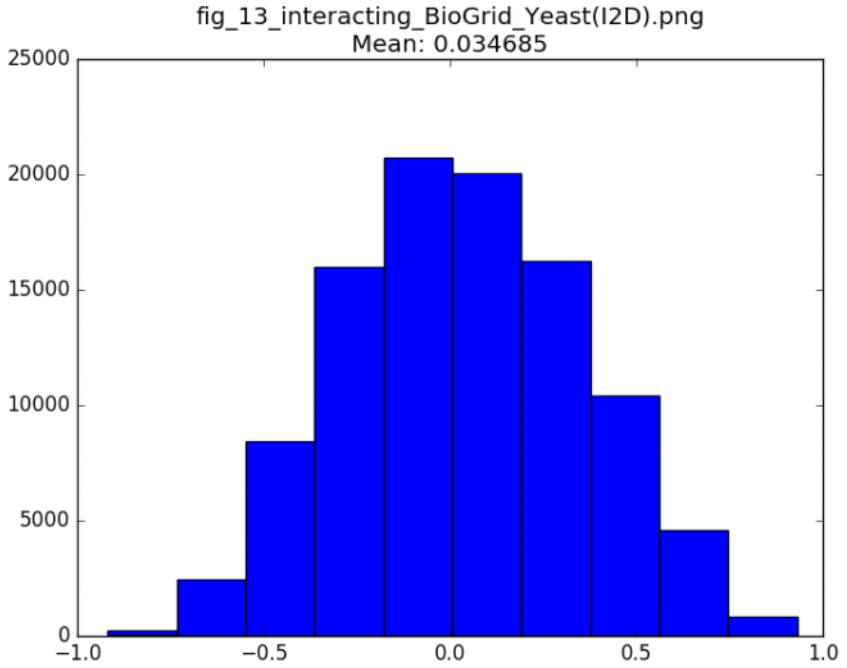


fig_13_interacting_BioGrid_Yeast(I2D).png
Mean: 0.034685
25000
20000
15000
10000
5000
0
−1.0
−0.5
0.0
0.5
1.0

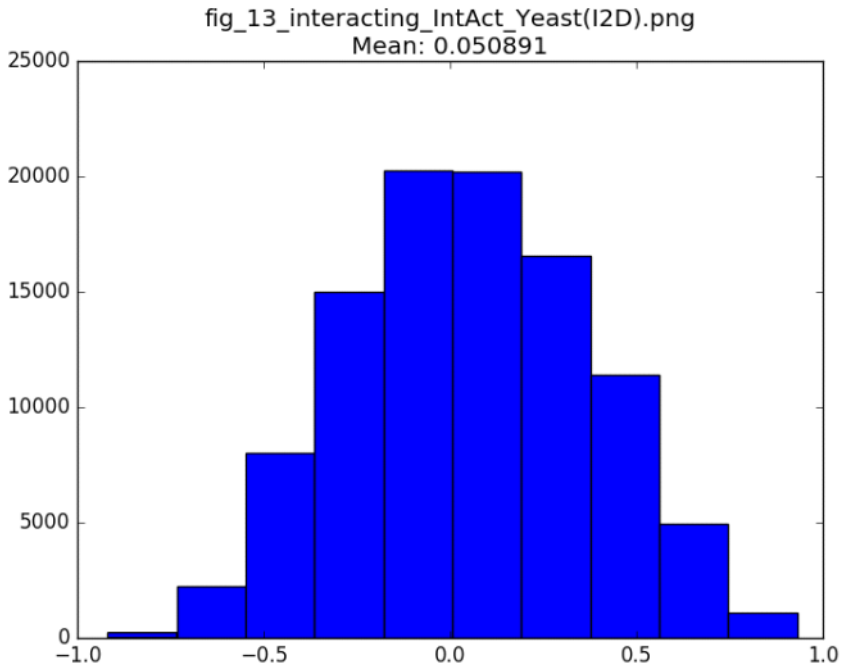


fig_13_interacting_IntAct_Yeast(I2D).png
Mean: 0.050891
25000
20000
15000
10000
5000
0
−1.0
−0.5
0.0
0.5
1.0

Interpretation/notes:

Tuesday, August 16, 2016 6:44 PM

Other notes: Could check whether the domain annotations from pfam give any signal vs Brain's annotations. Need to check better if the ref seq versions match the gene name -- in other words if my sequences and brian's sequences really are equivalent.

Inparanoid

Sunday, August 21, 2016 11:26 PM

Completed inparanoid runs:

Pan troglodytes
Pongo pygmaeus abelii (orangutan),
Macaca mulatta (rhesus macaque),
Callithrix jacchus (marmoset),
Microcebus murinus (mouse lemur),
Otolemur garnettii (bushbaby),
Tupaia belangeri (tree shrew),
Cavia porcellus (guinea pig),
Dipodomys ordii (kangaroo rat),
Mus musculus (mouse),
Rattus norvegicus (rat),
Oryctolagus cuniculus (rabbit),
Ochotona princeps (pika),
Vicugna pacos (alpaca),
Sorex araneus (shrew),
Bos taurus(cow),
Tursiops truncatus (dolphin),
Pteropus vampyrus (megabat),
Myotis lucifugus (microbat),
Erinaceus europaeus (hedgehog),
Equus caballus (horse),
Canis lupus familiaris (dog),
Felis catus (cat),
Echinops telfairi (tenrec),
Loxodonta africana(elephant),
Dasypus novemcinctus (armadillo),
-Monodelphis domestica (opossum),
Ornithorhynchus anatinus (platypus)

	Exit status = 0
—	Exit status = 271 (Possibly ran out of memory)
	Present in phylo tree in Nathan's paper
	Still running (9/1)
.	Have output for [18/28]

It looks like the ones that have output tended to have run for much fewer hours than the others. Looking at the file sizes of the succesful ones vs not successful, the successful ones are more likely to be smaller. Range is 12 - 61 mb, median 24 mb

Also have this error coming up (unable to get accounting entry for 94499.service1), but it's not present in all failures

Heat map outline

Tuesday, August 23, 2016 8:34 PM

Sliding windows will be 25 amino acids long, any remaining amino acids need to be 10 in length to be included (can't have tiny sequences going into paml)

Go through the selected genes (alignments) and create new alignment based on the sliding window I choose, everything else up to the ERC values should be the same:

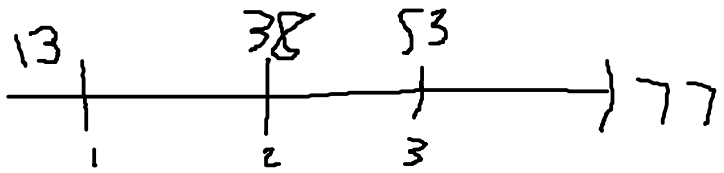
> New, shorter, alignments
> Create pfam files, etc.
> Run pfam to get trees with branch lengths
> Run ERC script … will need average tree data … but this might be able to be based on the domain level stuff?
> The average tree is a measure of the expected rates of evolution for each species. This may depend on sequence length, but since the sequences already being used (without the sliding window) vary in size, it *might* be ok. I'll go with it and see how it comes out.

Once I have the ERC values I can line them up with the sequences in a matrix, with each window centered on the amino acid in the middle (shift everything over half the window size).

Will need to produce heat map, should be a good way to do that in python,
maybe ggplot etc.

Will also need time to pick out some examples and analyze them, very important to be able to interpret them and see if they are significant.

Heat map windows compared to domain sequences for ERC script to produce average tree. Could be problematic?
Running



CDC14 vs NET1

Thursday, August 25, 2016 2:22 PM

CDC14
In S. cerevisiae, Cdc14 is regulated by its competitive inhibitor Cfi/Net1, which localizes Cdc14 to the nucleolus.[28] During anaphase, Cdc14 is "uncaged" and spreads to the rest of the cell. Two networks mediate the release of Cdc14 from the nucleolus: FEAR (CDC Fourteen Early Anaphase Release) and MEN (Mitotic Exit Network); while these networks are complex, it is thought that these networks result in the phosphorylation of Cfi/Net1 and/or Cdc14, resulting in disassociation of the complex. In S. pombe, phosphorylation of the Cdc14 ortholog by Cdk1 is known to directly inhibit the catalytic activity of the phosphatase.[29]

From <https://en.wikipedia.org/wiki/Cdc14>

Protein phosphatase which antagonizes mitotic cyclin-dependent kinase CDC28, the inactivation of which is essential for exit from mitosis. To access its substrates, is released from nucleolar sequestration during mitosis. Plays an essential in coordinating the nuclear division cycle with cytokinesis through the cytokinesis checkpoint. Involved in chromosome segregation, where it is required for meiosis I spindle dissambly as well as for establishing two consecutive chromosome segregation phases. Allows damaged actomyosin rings to be maintained to facilitate completion of cell division in response to minor perturbation of the cell division machinery. Inhibits transcription of ribosomal genes (rDNA) during anaphase and controls segregation of nucleolus by facilitating condensin targeting to rDNA chromatin in anaphase. Dephosphorylates SIC1, a CDC28 inhibitor, and SWI5, a transcription factor for SIC1, and induces degradation of mitotic cyclins, likely by dephosphorylating the activator of mitotic cyclin degradation, CDH1. Dephosphorylates the microtubule bundling factor ASE1 which is required to define a centered and focused mitotic spindle midzone that can drive continuous spindle elongation. Dephosphorylates the anaphase-promoting complex inhibitor ACM1, leading to its degradation. Facilitates INN1-CYK3 complex formation which promotes cytokinesis through the dephosphorylation of CDC28-phosphosphorylated INN1. Reverts also the inhibitory CDC28 phosphorylation of CHS2 for endoplasmic reticulum export, ensuring that septum formation is contingent upon chromosome separation and exit from mitosis. Additional substrates for CDC14 are the formins BNI1 and BNR1, as well as CDC6, DBP2, DSN1, INCENP, KAR9, MCM3, ORC2, ORC6, SLD2, and SWI6. Activity is inhibited by interaction with NET1 which sequesters it to the nucleolus.

From <http://www.uniprot.org/uniprot/Q00684>

NET1
Acts as guanine nucleotide exchange factor (GEF) for RhoA GTPase. May be involved in activation of the SAPK/JNK pathway Stimulates genotoxic stress-induced RHOB activity in breast cancer cells leading to their cell death.

From <http://www.uniprot.org/uniprot/Q7Z628>

Interpretation

Thursday, August 25, 2016 3:26 PM

Meeting Notes:

Brian domain references: PLOS genetics website

Deep Mutational Scanning

Wednesday, September 28, 2016 5:01 PM

413.full

Massively Parallel Functional Analysis of BRCA1 RING Domain Variants

From <http://www.genetics.org/content/200/2/413>

nmeth.3027

Deep mutational scanning: a new style of protein science

Stanley Fields

From <http://www.nature.com/nmeth/journal/v11/n8/full/nmeth.3027.html>

https://apps.webofknowledge.com/full_record.do?product=WOS&search_mode=GeneralSearch&qid=4&SID=4BfHdbLnRuLbzKIyI6d&page=1&doc=6

nmeth.1492

High-resolution mapping of protein sequence-function relationships

Stanley Fields

From <http://www.nature.com/nmeth/journal/v7/n9/full/nmeth.1492.html>
https://apps.webofknowledge.com/full_record.do?product=WOS&search_mode=GeneralSearch&qid=4&SID=4BfHdbLnRuLbzKIyI6d&page=1&doc=5

Genome

Genome Res.-2016...

An extended set of yeast-based functional assays accurately identifies human disease mutations

From <http://genome.cshlp.org/content/26/5/670>
https://apps.webofknowledge.com/full_record.do?product=WOS&search_mode=CitingArticles&qid=8&SID=4BfHdbLnRuLbzKIyI6d&page=1&doc=3&cacheurlFromRightClick=no

nprot.2012.069

Fitness analyses of all possible point mutations for regions of genes in yeast

From <http://www.nature.com/nprot/journal/v7/n7/full/nprot.2012.069.html>

https://apps.webofknowledge.com/full_record.do?product=WOS&search_mode=GeneralSearch&qid=29&SID=4BfHdbLnRuLbzKIyI6d&page=1&doc=1

High-resolution mapping of protein sequence-function relationships

Wednesday, September 28, 2016 5:17 PM

Relevant
Maybe
Not Sure
Not Relevant

Cited by 145:

FN Thomson Reuters Web of Science™
VR 1.0
PT J
AU Wu, NC
  Dai, L
  Olson, CA
  Lloyd-Smith, JO
  Sun, R
AF Wu, Nicholas C.
  Dai, Lei
  Olson, C. Anders
  Lloyd-Smith, James O.
  Sun, Ren
TI Adaptation in protein fitness landscapes is facilitated by indirect
  paths
SO eLife
AB The structure of fitness landscapes is critical for understanding adaptive protein evolution. Previous empirical studies on fitness landscapes were confined to either the neighborhood around the wild type sequence, involving mostly single and double mutants, or a combinatorially complete subgraph involving only two amino acids at each site. In reality, the dimensionality of protein sequence space is higher (20(L)) and there may be higher-order interactions among more than two sites. Here we experimentally characterized the fitness landscape of four sites in protein GB1, containing 20(4) = 160,000 variants. We found that while reciprocal sign epistasis blocked many direct paths of adaptation, such evolutionary traps could be circumvented by indirect paths through genotype space involving gain and subsequent loss of mutations. These indirect paths alleviate the constraint on adaptive protein evolution, suggesting that the heretofore neglected dimensions of sequence space may change our views on how proteins evolve.
SN 2050-084X
PD JUL 8
PY 2016
VL 5
UT WOS:000380855300001
ER

PT J
AU Nadler, DC
  Morgan, SA
  Flamholz, A
  Kortright, KE
  Savage, DF
AF Nadler, Dana C.
  Morgan, Stacy-Anne
  Flamholz, Avi
  Kortright, Kaitlyn E.
  Savage, David F.
TI Rapid construction of metabolite biosensors using domain-insertion
  profiling
SO NATURE COMMUNICATIONS
AB Single-fluorescent protein biosensors (SFPBs) are an important class of probes that enable the single-cell quantification of analytes in vivo. Despite advantages over other detection technologies, their use has been limited by the inherent challenges of their construction. Specifically, the rational design of green fluorescent protein (GFP) insertion into a ligand-binding domain, generating the requisite allosteric coupling, remains a rate-limiting step. Here, we describe an unbiased approach, termed domain-insertion profiling with DNA sequencing (DIP-seq), that combines the rapid creation of diverse libraries of potential SFPBs and high-throughput activity assays to identify functional biosensors. As a proof of concept, we construct an SFPB for the important regulatory sugar trehalose. DIP-seq analysis of a trehalose-binding-protein reveals allosteric hotspots for GFP insertion and results in high-dynamic range biosensors that function robustly in vivo. Taken together, DIP-seq simultaneously accelerates metabolite biosensor construction and provides a novel tool for interrogating protein allostery.
SN 2041-1723
PD JUL
PY 2016

VL 7
AR 12266
DI 10.1038/ncomms12266
UT WOS:000380535900001
PM 27470466
ER

PT J
AU Starr, TN
Thornton, JW
AF Starr, Tyler N.
Thornton, Joseph W.
TI Epistasis in protein evolution
SO PROTEIN SCIENCE
AB The structure, function, and evolution of proteins depend on physical and genetic interactions among amino acids. Recent studies have used new strategies to explore the prevalence, biochemical mechanisms, and evolutionary implications of these interactions-called epistasis-within proteins. Here we describe an emerging picture of pervasive epistasis in which the physical and biological effects of mutations change over the course of evolution in a lineage-specific fashion. Epistasis can restrict the trajectories available to an evolving protein or open new paths to sequences and functions that would otherwise have been inaccessible. We describe two broad classes of epistatic interactions, which arise from different physical mechanisms and have different effects on evolutionary processes. Specific epistasis-in which one mutation influences the phenotypic effect of few other mutations-is caused by direct and indirect physical interactions between mutations, which nonadditively change the protein's physical properties, such as conformation, stability, or affinity for ligands. In contrast, nonspecific epistasis describes mutations that modify the effect of many others; these typically behave additively with respect to the physical properties of a protein but exhibit epistasis because of a nonlinear relationship between the physical properties and their biological effects, such as function or fitness. Both types of interaction are rampant, but specific epistasis has stronger effects on the rate and outcomes of evolution, because it imposes stricter constraints and modulates evolutionary potential more dramatically; it therefore makes evolution more contingent on low-probability historical events and leaves stronger marks on the sequences, structures, and functions of protein families.
SN 0961-8368
EI 1469-896X
PD JUL
PY 2016
VL 25
IS 7
SI SI
BP 1204
EP 1218
DI 10.1002/pro.2897
UT WOS:000380067400005
PM 26833806
ER

PT J
AU Abriata, LA
Bovigny, C
Dal Peraro, M
AF Abriata, Luciano A.
Bovigny, Christophe
Dal Peraro, Matteo
TI Detection and sequence/structure mapping of biophysical constraints to protein variation in saturated mutational libraries and protein sequence alignments with a dedicated server
SO BMC BIOINFORMATICS
AB Background: Protein variability can now be studied by measuring high-resolution tolerance-to-substitution maps and fitness landscapes in saturated mutational libraries. But these rich and expensive datasets are typically interpreted coarsely, restricting detailed analyses to positions of extremely high or low variability or dubbed important beforehand based on existing knowledge about active sites, interaction surfaces, (de) stabilizing mutations, etc.
Results: Our new webserver PsychoProt (freely available without registration at http://psychoprot.epfl.ch or at http://lucianoabriata.altervista.org/psychoprot/index.html) helps to detect, quantify, and sequence/structure map the biophysical and biochemical traits that shape amino acid preferences throughout a protein as determined by deep-sequencing of saturated mutational libraries or from large alignments of naturally occurring variants.
Discussion: We exemplify how PsychoProt helps to (i) unveil protein structure-function relationships from experiments and from alignments that are consistent with structures according to coevolution analysis, (ii) recall global information about structural and functional features and identify hitherto unknown constraints to variation in alignments, and (iii) point at different sources of variation among related experimental datasets or between experimental and alignment-based data. Remarkably, metabolic costs of the amino acids pose strong constraints to variability at protein surfaces in nature but not in the laboratory. This and other differences call for caution when extrapolating results from in vitro experiments to natural scenarios in, for example, studies of protein evolution.
Conclusion: We show through examples how PsychoProt can be a useful tool for the broad communities of structural biology and molecular evolution, particularly for studies about protein modeling, evolution and design.
SN 1471-2105

PD JUN 17
PY 2016
VL 17
AR 242
DI 10.1186/s12859-016-1124-4
UT WOS:000378845300001
PM 27315797
ER

PT J
AU Shendure, J
Fields, S
AF Shendure, Jay
Fields, Stanley
TI Massively Parallel Genetics
SO GENETICS
AB Human genetics has historically depended on the identification of individuals whose natural genetic variation underlies an observable trait or disease risk. Here we argue that new technologies now augment this historical approach by allowing the use of massively parallel assays in model systems to measure the functional effects of genetic variation in many human genes. These studies will help establish the disease risk of both observed and potential genetic variants and to overcome the problem of "variants of uncertain significance."
SN 0016-6731
EI 1943-2631
PD JUN
PY 2016
VL 203
IS 2
BP 617
EP 619
DI 10.1534/genetics.115.180562
UT WOS:000377462800003
PM 27270695
ER

PT J
AU Sarkisyan, KS
Bolotin, DA
Meer, MV
Usmanova, DR
Mishin, AS
Sharonov, GV
Ivankov, DN
Bozhanova, NG
Baranov, MS
Soylemez, O
Bogatyreva, NS
Vlasov, PK
Egorov, ES
Logacheva, MD
Kondrashov, AS
Chudakov, DM
Putintseva, EV
Mamedov, IZ
Tawfik, DS
Lukyanov, KA
Kondrashov, FA
AF Sarkisyan, Karen S.
Bolotin, Dmitry A.
Meer, Margarita V.
Usmanova, Dinara R.
Mishin, Alexander S.
Sharonov, George V.
Ivankov, Dmitry N.
Bozhanova, Nina G.
Baranov, Mikhail S.
Soylemez, Onuralp
Bogatyreva, Natalya S.
Vlasov, Peter K.

Egorov, Evgeny S.
Logacheva, Maria D.
Kondrashov, Alexey S.
Chudakov, Dmitry M.
Putintseva, Ekaterina V.
Mamedov, Ilgar Z.
Tawfik, Dan S.
Lukyanov, Konstantin A.
Kondrashov, Fyodor A.
TI Local fitness landscape of the green fluorescent protein
SO NATURE
AB Fitness landscapes(1,2) depict how genotypes manifest at the phenotypic level and form the basis of our understanding of many areas of biology(2-7), yet their properties remain elusive. Previous studies have analysed specific genes, often using their function as a proxy for fitness(2,4), experimentally assessing the effect on function of single mutations and their combinations in a specific sequence(2,5,8-15) or in different sequences(2,3,5,16-18). However, systematic high-throughput studies of the local fitness landscape of an entire protein have not yet been reported. Here we visualize an extensive region of the local fitness landscape of the green fluorescent protein from Aequorea victoria (avGFP) by measuring the native function (fluorescence) of tens of thousands of derivative genotypes of avGFP. We show that the fitness landscape of avGFP is narrow, with 3/4 of the derivatives with a single mutation showing reduced fluorescence and half of the derivatives with four mutations being completely non-fluorescent. The narrowness is enhanced by epistasis, which was detected in up to 30% of genotypes with multiple mutations and mostly occurred through the cumulative effect of slightly deleterious mutations causing a threshold-like decrease in protein stability and a concomitant loss of fluorescence. A model of orthologous sequence divergence spanning hundreds of millions of years predicted the extent of epistasis in our data, indicating congruence between the fitness landscape properties at the local and global scales. The characterization of the local fitness landscape of avGFP has important implications for several fields including molecular evolution, population genetics and protein design.
RI Mishin, Alexander/L-9420-2014; Baranov, Mikhail/L-5014-2016; Logacheva, Maria/K-5217-2012
OI Mishin, Alexander/0000-0002-4935-7030;
SN 0028-0836
EI 1476-4687
PD MAY 19
PY 2016
VL 533
IS 7603
BP 397
EP +
DI 10.1038/nature17995
UT WOS:000376004300049
PM 27193686
ER

PT J
AU Boyer, S
Biswas, D
Soshee, AK
Scaramozzino, N
Nizak, C
Rivoire, O
AF Boyer, Sebastien
Biswas, Dipanwita
Soshee, Ananda Kumar
Scaramozzino, Natale
Nizak, Clement
Rivoire, Olivier
TI Hierarchy and extremes in selections from pools of randomized proteins
SO PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA
AB Variation and selection are the core principles of Darwinian evolution, but quantitatively relating the diversity of a population to its capacity to respond to selection is challenging. Here, we examine this problem at a molecular level in the context of populations of partially randomized proteins selected for binding to well-defined targets. We built severalminimal protein libraries, screened themin vitro by phage display, and analyzed their response to selection by high-throughput sequencing. A statistical analysis of the results reveals two main findings. First, libraries with the same sequence diversity but built around different "frameworks" typically have vastly different responses; second, the distribution of responses of the best binders in a library follows a simple scaling law. We show how an elementary probabilistic model based on extreme value theory rationalizes the latter finding. Our results have implications for designing synthetic protein libraries, estimating the density of functional biomolecules in sequence space, characterizing diversity in natural populations, and experimentally investigating evolvability (i.e., the potential for future evolution).
RI Nizak, Clement/L-1334-2013
OI Nizak, Clement/0000-0002-4591-8240
SN 0027-8424
PD MAR 29

PY 2016
VL 113
IS 13
BP 3482
EP 3487
DI 10.1073/pnas.1517813113
UT WOS:000372876400041
PM 26969726
ER

PT J
AU Peterman, N
Levine, E
AF Peterman, Neil
Levine, Erel
TI Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations
SO BMC GENOMICS
AB Background: Sort-seq is an effective approach for simultaneous activity measurements in a large-scale library, combining flow cytometry, deep sequencing, and statistical inference. Such assays enable the characterization of functional landscapes at unprecedented scale for a wide-reaching array of biological molecules and functionalities in vivo. Applications of sort-seq range from footprinting to establishing quantitative models of biological systems and rational design of synthetic genetic elements. Nearly as diverse are implementations of this technique, reflecting key design choices with extensive impact on the scope and accuracy the results. Yet how to make these choices remains unclear. Here we investigate the effects of alternative sort-seq designs and inference methods on the information output using mathematical formulation and simulations.
Results: We identify key intrinsic properties of any system of interest with practical implications for sort-seq assays, depending on the experimental goals. The fluorescence range and cell-to-cell variability specify the number of sorted populations needed for quantitative measurements that are precise and unbiased. These factors also indicate cases where an enrichment-based approach that uses a single sorted population can offer satisfactory results. These predications of our model are corroborated using re-analysis of published data. We explore implications of these results for quantitative modeling and library design.
Conclusions: Sort-seq assays can be streamlined by reducing the number of sorted populations, saving considerable resources. Simple preliminary experiments can guide optimal experiment design, minimizing cost while maintaining the maximal information output and avoiding latent biases. These insights can facilitate future applications of this highly adaptable technique.
SN 1471-2164
PD MAR 9
PY 2016
VL 17
AR 206
DI 10.1186/s12864-016-2533-5
UT WOS:000371595800001
PM 26956374
ER

PT J
AU Wilburn, DB
Swanson, WJ
AF Wilburn, Damien B.
Swanson, Willie J.
TI From molecules to mating: Rapid evolution and biochemical studies of reproductive proteins
SO JOURNAL OF PROTEOMICS
AB Sexual reproduction and the exchange of genetic information are essential biological processes for species across all branches of the tree of life. Over the last four decades, biochemists have continued to identify many of the factors that facilitate reproduction, but the molecular mechanisms that mediate this process continue to elude us. However, a recurring observation in this research has been the rapid evolution of reproductive proteins. In animals, the competing interests of males and females often result in arms race dynamics between pairs of interacting proteins. This phenomenon has been observed in all stages of reproduction, including pheromones, seminal fluid components, and gamete recognition proteins. In this article, we review how the integration of evolutionary theory with biochemical experiments can be used to study interacting reproductive proteins. Examples are included from both model and non-model organisms, and recent studies are highlighted for their use of stateof-the-art genomic and proteomic techniques.
Significance: Despite decades of research, our understanding of the molecular mechanisms that mediate fertilization remain poorly characterized. To date, molecular evolutionary studies on both model and non-model organisms have provided some of the best inferences to elucidating the molecular underpinnings of animal reproduction. This review article details how biochemical and evolutionary experiments have jointly enhanced the field for 40 years, and how recent work using high-throughput genomic and proteomic techniques have shed additional insights into this crucial biological process.

SN 1874-3919
EI 1876-7737
PD MAR 1
PY 2016
VL 135

SI SI
BP 12
EP 25
DI 10.1016/j.jprot.2015.06.007
UT WOS:000372685800003
PM 26074353
ER

PT J
AU Atwal, GS
Kinney, JB
AF Atwal, Gurinder S.
Kinney, Justin B.
TI Learning Quantitative Sequence-Function Relationships from Massively Parallel Experiments
SO JOURNAL OF STATISTICAL PHYSICS
AB A fundamental aspect of biological information processing is the ubiquity of sequence-function relationships-functions that map the sequence of DNA, RNA, or protein to a biochemically relevant activity. Most sequence-function relationships in biology are quantitative, but only recently have experimental techniques for effectively measuring these relationships been developed. The advent of such "massively parallel" experiments presents an exciting opportunity for the concepts and methods of statistical physics to inform the study of biological systems. After reviewing these recent experimental advances, we focus on the problem of how to infer parametric models of sequence-function relationships from the data produced by these experiments. Specifically, we retrace and extend recent theoretical work showing that inference based on mutual information, not the standard likelihood-based approach, is often necessary for accurately learning the parameters of these models. Closely connected with this result is the emergence of "diffeomorphic modes"aEuro"directions in parameter space that are far less constrained by data than likelihood-based inference would suggest. Analogous to Goldstone modes in physics, diffeomorphic modes arise from an arbitrarily broken symmetry of the inference problem. An analytically tractable model of a massively parallel experiment is then described, providing an explicit demonstration of these fundamental aspects of statistical inference. This paper concludes with an outlook on the theoretical and computational challenges currently facing studies of quantitative sequence-function relationships.
OI Kinney, Justin/0000-0003-1897-3778
SN 0022-4715
EI 1572-9613
PD MAR
PY 2016
VL 162
IS 5
SI SI
BP 1203
EP 1243
DI 10.1007/s10955-015-1398-3
UT WOS:000371088000007
ER

PT J
AU Phillips, AM
Shoulders, MD
AF Phillips, Angela M.
Shoulders, Matthew D.
TI The Path of Least Resistance: Mechanisms to Reduce Influenza's Sensitivity to Oseltamivir
SO JOURNAL OF MOLECULAR BIOLOGY
SN 0022-2836
EI 1089-8638
PD FEB 13
PY 2016
VL 428
IS 3
BP 533
EP 537
DI 10.1016/j.jmb.2015.12.019
UT WOS:000371839800001
PM 26748011
ER

PT J
AU Jiang, L
Liu, P
Bank, C

Renzette, N
Prachanronarong, K
Yilmaz, LS
Caffrey, DR
Zeldovich, KB
Schiffer, CA
Kowalik, TF
Jensen, JD
Finberg, RW
Wang, JP
Bolon, DNA
AF Jiang, Li
Liu, Ping
Bank, Claudia
Renzette, Nicholas
Prachanronarong, Kristina
Yilmaz, Lutfu S.
Caffrey, Daniel R.
Zeldovich, Konstantin B.
Schiffer, Celia A.
Kowalik, Timothy F.
Jensen, Jeffrey D.
Finberg, Robert W.
Wang, Jennifer P.
Bolon, Daniel N. A.
TI A Balance between Inhibitor Binding and Substrate Processing Confers Influenza Drug Resistance
SO JOURNAL OF MOLECULAR BIOLOGY
AB The therapeutic benefits of the neuraminidase (NA) inhibitor oseltamivir are dampened by the emergence of drug resistance mutations in influenza A virus (IAV). To investigate the mechanistic features that underlie resistance, we developed an approach to quantify the effects of all possible single-nucleotide substitutions introduced into important regions of NA. We determined the experimental fitness effects of 450 nucleotide mutations encoding positions both surrounding the active site and at more distant sites in an N1 strain of IAV in the presence and absence of oseltamivir. NA mutations previously known to confer oseltamivir resistance in N1 strains, including H275Y and N295S, were adaptive in the presence of drug, indicating that our experimental system captured salient features of real-world selection pressures acting on NA. We identified mutations, including several at position 223, that reduce the apparent affinity for oseltamivir in vitro. Position 223 of NA is located adjacent to a hydrophobic portion of oseltamivir that is chemically distinct from the substrate, making it a hotspot for substitutions that preferentially impact drug binding relative to substrate processing. Furthermore, two NA mutations, K221N and Y276F, each reduce susceptibility to oseltamivir by increasing NA activity without altering drug binding. These results indicate that competitive expansion of IAV in the face of drug pressure is mediated by a balance between inhibitor binding and substrate processing.
SN 0022-2836
EI 1089-8638
PD FEB 13
PY 2016
VL 428
IS 3
BP 538
EP 553
DI 10.1016/j.jmb.2015.11.027
UT WOS:000371839800002
PM 26656922
ER

PT J
AU Zhang, TH
Wu, NC
Sun, R
AF Zhang, Tian-Hao
Wu, Nicholas C.
Sun, Ren
TI A benchmark study on error-correction by read-pairing and tag-clustering in amplicon-based deep sequencing
SO BMC GENOMICS
AB Background: The high error rate of next generation sequencing (NGS) restricts some of its applications, such as monitoring virus mutations and detecting rare mutations in tumors. There are two commonly employed sequencing library preparation strategies to improve sequencing accuracy by correcting sequencing errors: read-pairing method and tag-clustering method (i.e. primer ID or UID). Here, we constructed a homogeneous library from a single clone, and compared the variant calling accuracy of these error-correction methods.
Result: We comprehensively described the strengths and pitfalls of these methods. We found that both read-pairing and tag-clustering methods

significantly decreased sequencing error rate. While the read-pairing method was more effective than the tag-clustering method at correcting insertion and deletion errors, it was not as effective as the tag-clustering method at correcting substitution errors. In addition, we observed that when the read quality was poor, the tag-clustering method led to huge coverage loss. We also tested the effect of applying quality score filtering to the error-correction methods and demonstrated that quality score filtering was able to impose a minor, yet statistically significant improvement to the error-correction methods tested in this study.

Conclusion: Our study provides a benchmark for researchers to select suitable error-correction methods based on the goal of the experiment by balancing the trade-off between sequencing cost (i.e. sequencing coverage requirement) and detection sensitivity.

RI Wu, Nicholas/H-3822-2015
OI Wu, Nicholas/0000-0002-9078-6697
SN 1471-2164
PD FEB 12
PY 2016
VL 17
AR 108
DI 10.1186/s12864-016-2388-9
UT WOS:000370015400001
PM 26868371
ER

PT J
AU Fischer, M
Kang, M
Brindle, NP
AF Fischer, Marlies
Kang, Mandeep
Brindle, Nicholas Pj
TI Using experimental evolution to probe molecular mechanisms of protein function
SO PROTEIN SCIENCE
AB Directed evolution is a powerful tool for engineering protein function. The process of directed evolution involves iterative rounds of sequence diversification followed by assaying activity of variants and selection. The range of sequence variants and linked activities generated in the course of an evolution are a rich information source for investigating relationships between sequence and function. Key residue positions determining protein function, combinatorial contributors to activity and even potential functional mechanisms have been revealed in directed evolutions. The recent application of high throughput sequencing substantially increases the information that can be retrieved from directed evolution experiments. Combined with computational analysis this additional sequence information has allowed high-resolution analysis of individual residue contributions to activity. These developments promise to significantly enhance the depth of insight that experimental evolution provides into mechanisms of protein function.
SN 0961-8368
EI 1469-896X
PD FEB
PY 2016
VL 25
IS 2
BP 352
EP 359
DI 10.1002/pro.2836
UT WOS:000369819500005
PM 26509591
ER

PT J
AU Jin, Z
Di Rienzi, SC
Janzon, A
Werner, JJ
Angenent, LT
Dangl, JL
Fowler, DM
Ley, RE
AF Jin, Zhao
Di Rienzi, Sara C.
Janzon, Anders
Werner, Jeff J.
Angenent, Largus T.
Dangl, Jeffrey L.
Fowler, Douglas M.
Ley, Ruth E.
TI Novel Rhizosphere Soil Alleles for the Enzyme

Fowler, Douglas M.
Ley, Ruth E.
TI Novel Rhizosphere Soil Alleles for the Enzyme 1-Aminocyclopropane-1-Carboxylate Deaminase Queried for Function with an In Vivo Competition Assay
SO APPLIED AND ENVIRONMENTAL MICROBIOLOGY
AB Metagenomes derived from environmental microbiota encode a vast diversity of protein homologs. How this diversity impacts protein function can be explored through selection assays aimed to optimize function. While artificially generated gene sequence pools are typically used in selection assays, their usage may be limited because of technical or ethical reasons. Here, we investigate an alternative strategy, the use of soil microbial DNA as a starting point. We demonstrate this approach by optimizing the function of a widely occurring soil bacterial enzyme, 1-aminocyclopropane-1-carboxylate (ACC) deaminase. We identified a specific ACC deaminase domain region (ACCD-DR) that, when PCR amplified from the soil, produced a variant pool that we could swap into functional plasmids carrying ACC deaminase-encoding genes. Functional clones of ACC deaminase were selected for in a competition assay based on their capacity to provide nitrogen to Escherichia coli in vitro. The most successful ACCD-DR variants were identified after multiple rounds of selection by sequence analysis. We observed that previously identified essential active-site residues were fixed in the original unselected library and that additional residues went to fixation after selection. We identified a divergent essential residue whose presence hints at the possible use of alternative substrates and a cluster of neutral residues that did not influence ACCD performance. Using an artificial ACCD-DR variant library generated by DNA oligomer synthesis, we validated the same fixation patterns. Our study demonstrates that soil metagenomes are useful starting pools of protein-coding-gene diversity that can be utilized for protein optimization and functional characterization when synthetic libraries are not appropriate.
SN 0099-2240
EI 1098-5336
PD FEB
PY 2016
VL 82
IS 4
BP 1050
EP 1059
DI 10.1128/AEM.03074-15
UT WOS:000369375900008
PM 26637602
ER

PT J
AU Echave, J
Spielman, SJ
Wilke, CO
AF Echave, Julian
Spielman, Stephanie J.
Wilke, Claus O.
TI Causes of evolutionary rate variation among protein sites
SO NATURE REVIEWS GENETICS
AB It has long been recognized that certain sites within a protein, such as sites in the protein core or catalytic residues in enzymes, are evolutionarily more conserved than other sites. However, our understanding of rate variation among sites remains surprisingly limited. Recent progress to address this includes the development of a wide array of reliable methods to estimate site-specific substitution rates from sequence alignments. In addition, several molecular traits have been identified that correlate with site-specific mutation rates, and novel mechanistic biophysical models have been proposed to explain the observed correlations. Nonetheless, current models explain, at best, approximately 60% of the observed variance, highlighting the limitations of current methods and models and the need for new research directions.
SN 1471-0056
EI 1471-0064
PD FEB
PY 2016
VL 17
IS 2
BP 109
EP 121
DI 10.1038/nrg.2015.18
UT WOS:000369179300012
PM 26781812
ER

PT J
AU Au, L
Green, DF
AF Au, Loretta
Green, David F.
TI Direct Calculation of Protein Fitness Landscapes through Computational Protein Design
SO BIOPHYSICAL JOURNAL

SO BIOPHYSICAL JOURNAL
AB Naturally selected amino-acid sequences or experimentally derived ones are often the basis for understanding how protein three-dimensional conformation and function are determined by primary structure. Such sequences for a protein family comprise only a small fraction of all possible variants, however, representing the fitness landscape with limited scope. Explicitly sampling and characterizing alternative, unexplored protein sequences would directly identify fundamental reasons for sequence robustness (or variability), and we demonstrate that computational methods offer an efficient mechanism toward this end, on a large scale. The dead-end elimination and A* search algorithms were used here to find all low-energy single mutant variants, and corresponding structures of a G-protein heterotrimer, to measure changes in structural stability and binding interactions to define a protein fitness landscape. We established consistency between these algorithms with known biophysical and evolutionary trends for amino-acid substitutions, and could thus recapitulate known protein side-chain interactions and predict novel ones.
SN 0006-3495
EI 1542-0086
PD JAN 5
PY 2016
VL 110
IS 1
BP 75
EP 84
DI 10.1016/j.bpj.2015.11.029
UT WOS:000367783900030
PM 26745411
ER

PT J
AU Sun, DD
Xu, CR
Zhang, YS
AF Sun, Dandan
Xu, Chunrui
Zhang, Yusen
TI A Novel Method of 2D Graphical Representation for Proteins and Its Application
SO MATCH-COMMUNICATIONS IN MATHEMATICAL AND IN COMPUTER CHEMISTRY
AB In this paper, we propose the graph energy of 20 amino acids and the 2D graphical representation of protein sequences based on six physicochemical properties of 20 amino acids and the relationship between them. Moreover, we could get a specific vector from the graphical curve of a protein sequence, and use this vector to calculate the distance between two sequences. This approach avoids considering the differences in length of protein sequences. Finally, we research the similarities/dissimilarities of ND5 and 36PDs using our method and get better results compared with ClustalX2.
SN 0340-6253
PY 2016
VL 75
IS 2
BP 431
EP 446
UT WOS:000374814100016
ER

PT J
AU Leung, MKK
Delong, A
Alipanahi, B
Frey, BJ
AF Leung, Michael K. K.
Delong, Andrew
Alipanahi, Babak
Frey, Brendan J.
TI Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets
SO PROCEEDINGS OF THE IEEE
AB In this paper, we provide an introduction to machine learning tasks that address important problems in genomic medicine. One of the goals of genomic medicine is to determine how variations in the DNA of individuals can affect the risk of different diseases, and to find causal explanations so that targeted therapies can be designed. Here we focus on how machine learning can help to model the relationship between DNA and the quantities of key molecules in the cell, with the premise that these quantities, which we refer to as cell variables, may be associated with disease risks. Modern biology allows high-throughput measurement of many such cell variables, including gene expression, splicing, and proteins binding to nucleic acids, which can all be treated as training targets for predictive models. With the growing availability of large-scale data sets and advanced computational techniques such as deep learning, researchers can help to usher in a new era of effective genomic medicine.
OI Alipanahi, Babak/0000-0001-8216-7178
SN 0018-9219
EI 1558-2256

PD JAN
PY 2016
VL 104
IS 1
SI SI
BP 176
EP 197
DI 10.1109/JPROC.2015.2494198
UT WOS:000367250300011
ER

PT J
AU Sahoo, A
Khare, S
Devanarayanan, S
Jain, PC
Varadarajan, R
AF Sahoo, Anusmita
Khare, Shruti
Devanarayanan, Sivasankar
Jain, Pankaj C.
Varadarajan, Raghavan
TI Residue proximity information and protein model discrimination using saturation-suppressor mutagenesis
SO ELIFE
AB Identification of residue-residue contacts from primary sequence can be used to guide protein structure prediction. Using Escherichia coli CcdB as the test case, we describe an experimental method termed saturation-suppressor mutagenesis to acquire residue contact information. In this methodology, for each of five inactive CcdB mutants, exhaustive screens for suppressors were performed. Proximal suppressors were accurately discriminated from distal suppressors based on their phenotypes when present as single mutants. Experimentally identified putative proximal pairs formed spatial constraints to recover >98% of native-like models of CcdB from a decoy dataset. Suppressor methodology was also applied to the integral membrane protein, diacylglycerol kinase A where the structures determined by X-ray crystallography and NMR were significantly different. Suppressor as well as sequence co-variation data clearly point to the Xray structure being the functional one adopted in vivo. The methodology is applicable to any macromolecular system for which a convenient phenotypic assay exists.
SN 2050-084X
PD DEC 30
PY 2015
VL 4
AR e09532
DI 10.7554/eLife.09532
UT WOS:000373812600001
ER

PT J
AU Nomine, Y
Choulier, L
Trave, G
Vernet, T
Altschuh, D
AF Nomine, Yves
Choulier, Laurence
Trave, Gilles
Vernet, Thierry
Altschuh, Daniele
TI Antibody Binding Selectivity: Alternative Sets of Antigen Residues Entail High-Affinity Recognition
SO PLOS ONE
AB Understanding the relationship between protein sequence and molecular recognition selectivity remains a major challenge. The antibody fragment scFv1F4 recognizes with sub nM affinity a decapeptide (sequence (6)TAMFQDPQER(15)) derived from the N-terminal end of human papilloma virus E6 oncoprotein. Using this decapeptide as antigen, we had previously shown that only the wild type amino-acid or conservative replacements were allowed at positions 9 to 12 and 15 of the peptide, indicating a strong binding selectivity. Nevertheless phenylalanine (F) was equally well tolerated as the wild type glutamine (Q) at position 13, while all other amino acids led to weaker scFv binding. The interfaces of complexes involving either Q or F are expected to diverge, due to the different physico-chemistry of these residues. This would imply that high-affinity binding can be achieved through distinct interfacial geometries. In order to investigate this point, we disrupted the scFv-peptide interface by modifying one or several peptide positions. We then analyzed the effect on binding of amino acid changes at the remaining positions, an altered susceptibility being indicative of an altered role in complex formation. The 23 starting variants analyzed contained replacements whose effects on scFv1F4 binding ranged from minor to drastic. A permutation analysis (effect of replacing each peptide position by all other amino acids except cysteine) was carried out on the 23 variants using the PEPperCHIP (R) Platform technology. A comparison of their permutation patterns with that of the wild type peptide indicated that starting replacements at position 11,

12 or 13 modified the tolerance to amino-acid changes at the other two positions. The interdependence between the three positions was confirmed by SPR (Biacore (R) technology). Our data demonstrate that binding selectivity does not preclude the existence of alternative high-affinity recognition modes.
SN 1932-6203
PD DEC 2
PY 2015
VL 10
IS 12
AR e0143374
DI 10.1371/journal.pone.0143374
UT WOS:000365926300044
PM 26629896
ER

PT J
AU Kelly, JT
De Colibus, L
Elliott, L
Fry, EE
Stuart, DI
Rowlands, DJ
Stonehouse, NJ
AF Kelly, James T.
De Colibus, Luigi
Elliott, Lauren
Fry, Elizabeth E.
Stuart, David I.
Rowlands, David J.
Stonehouse, Nicola J.
TI Potent antiviral agents fail to elicit genetically-stable resistance mutations in either enterovirus 71 or Coxsackievirus A16
SO ANTIVIRAL RESEARCH
AB Enterovirus 71 (EV71) and Coxsackievirus A16 (CVA16) are the two major causative agents of hand, foot and mouth disease (HFMD), for which there are currently no licenced treatments. Here, the acquisition of resistance towards two novel capsid-binding compounds, NLD and ALD, was studied and compared to the analogous compound GPP3. During serial passage, EV71 rapidly became resistant to each compound and mutations at residues I113 and V123 in VP1 were identified. A mutation at residue 113 was also identified in CVA16 after passage with GPP3. The mutations were associated with reduced thermostability and were rapidly lost in the absence of inhibitors. In silico modelling suggested that the mutations prevented the compounds from binding the VP1 pocket in the capsid. Although both viruses developed resistance to these potent pocket-binding compounds, the acquired mutations were associated with large fitness costs and reverted to WT phenotype and sequence rapidly in the absence of inhibitors. The most effective inhibitor, NLD, had a very large selectivity index, showing interesting pharmacological properties as a novel anti-EV71 agent. (C) 2015 The Authors. Published by Elsevier B.V.
SN 0166-3542
EI 1872-9096
PD DEC
PY 2015
VL 124
BP 77
EP 82
DI 10.1016/j.antiviral.2015.10.006
UT WOS:000370463200009
PM 26522770
ER

PT J
AU Bar-Shira, O
Maor, R
Chechik, G
AF Bar-Shira, Ossnat
Maor, Ronnie
Chechik, Gal
TI Gene Expression Switching of Receptor Subunits in Human Brain Development
SO PLOS COMPUTATIONAL BIOLOGY
AB Synaptic receptors in the human brain consist of multiple protein subunits, many of which have multiple variants, coded by different genes, and are differentially expressed across brain regions and developmental stages. The brain can tune the electrophysiological properties of synapses to regulate plasticity and information processing by switching from one protein variant to another. Such condition-dependent variant switch during development has been demonstrated in several neurotransmitter systems including NMDA and GABA. Here we systematically detect pairs of receptor-subunit variants

that switch during the lifetime of the human brain by analyzing postmortem expression data collected in a population of donors at various ages and brain regions measured using microarray and RNA-seq. To further detect variant pairs that co-vary across subjects, we present a method to quantify age-corrected expression correlation in face of strong temporal trends. This is achieved by computing the correlations in the residual expression beyond a cubic-spline model of the population temporal trend, and can be seen as a nonlinear version of partial correlations. Using these methods, we detect multiple new pairs of context dependent variants. For instance, we find a switch from GLRA2 to GLRA3 that differs from the known switch in the rat. We also detect an early switch from HTR1A to HTR5A whose trends are negatively correlated and find that their age-corrected expression is strongly positively correlated. Finally, we observe that GRIN2B switch to GRIN2A occurs mostly during embryonic development, presumably earlier than observed in rodents. These results provide a systematic map of developmental switching in the neurotransmitter systems of the human brain.
SN 1553-734X
EI 1553-7358
PD DEC
PY 2015
VL 11
IS 12
AR e1004559
DI 10.1371/journal.pcbi.1004559
UT WOS:000368521900009
PM 26636753
ER

PT J
AU Klesmith, JR
Bacik, JP
Michalczyk, R
Whitehead, TA
AF Klesmith, Justin R.
Bacik, John-Paul
Michalczyk, Ryszard
Whitehead, Timothy A.
TI Comprehensive Sequence-Flux Mapping of a Levoglucosan Utilization Pathway in E. coli
SO ACS SYNTHETIC BIOLOGY
AB Synthetic metabolic pathways often suffer from low specific productivity, and new methods that quickly assess pathway functionality for many thousands of variants are urgently needed. Here we present an approach that enables the rapid and parallel determination of sequence effects on flux for complete gene-encoding sequences. We show that this method can be used to determine the effects of over 8000 single point mutants of a pyrolysis oil catabolic pathway implanted in Escherichia coli. Experimental sequence-function data sets predicted whether fitness-enhancing mutations to the enzyme levoglucosan kinase resulted from enhanced catalytic efficiency or enzyme stability. A structure of one design incorporating 38 mutations elucidated the structural basis of high fitness mutations. One design incorporating 15 beneficial mutations supported a 15-fold improvement in growth rate and greater than 24-fold improvement in enzyme activity relative to the starting pathway. This technique can be extended to improve a wide variety of designed pathways.
OI Klesmith, Justin/0000-0003-2908-9355
SN 2161-5063
PD NOV
PY 2015
VL 4
IS 11
BP 1235
EP 1243
DI 10.1021/acssynbio.5b00131
UT WOS:000365461200009
PM 26369947
ER

PT J
AU Doud, MB
Ashenberg, O
Bloom, JD
AF Doud, Michael B.
Ashenberg, Orr
Bloom, Jesse D.
TI Site-Specific Amino Acid Preferences Are Mostly Conserved in Two Closely Related Protein Homologs
SO MOLECULAR BIOLOGY AND EVOLUTION
AB Evolution drives changes in a protein's sequence over time. The extent to which these changes in sequence lead to shifts in the underlying preference for each amino acid at each site is an important question with implications for comparative sequence-analysis methods, such as molecular phylogenetics. To quantify the extent that site-specific amino acid preferences shift during evolution, we performed deep mutational scanning on two homologs of human influenza nucleoprotein with 94% amino acid identity. We found that only a modest fraction of sites exhibited shifts in amino acid preferences

that exceeded the noise in our experiments. Furthermore, even among sites that did exhibit detectable shifts, the magnitude tended to be small relative to differences between nonhomologous proteins. Given the limited change in amino acid preferences between these close homologs, we tested whether our measurements could inform site-specific substitution models that describe the evolution of nucleoproteins from more diverse influenza viruses. We found that site-specific evolutionary models informed by our experiments greatly outperformed nonsite-specific alternatives in fitting phylogenies of nucleoproteins from human, swine, equine, and avian influenza. Combining the experimental data from both homologs improved phylogenetic fit, partly because measurements in multiple genetic contexts better captured the evolutionary average of the amino acid preferences for sites with shifting preferences. Our results show that site-specific amino acid preferences are sufficiently conserved that measuring mutational effects in one protein provides information that can improve quantitative evolutionary modeling of nearby homologs.
RI Bloom, Jesse/C-6837-2013
OI Bloom, Jesse/0000-0003-1267-3408
SN 0737-4038
EI 1537-1719
PD NOV
PY 2015
VL 32
IS 11
BP 2944
EP 2960
DI 10.1093/molbev/msv167
UT WOS:000363033100012
PM 26226986
ER

PT J
AU Bazzoli, A
Kelow, SP
Karanicolas, J
AF Bazzoli, Andrea
Kelow, Simon P.
Karanicolas, John
TI Enhancements to the Rosetta Energy Function Enable Improved Identification of Small Molecules that Inhibit Protein-Protein Interactions
SO PLOS ONE
AB Protein-protein interactions are among today's most exciting and promising targets for therapeutic intervention. To date, identifying small-molecules that selectively disrupt these interactions has proven particularly challenging for virtual screening tools, since these have typically been optimized to perform well on more "traditional" drug discovery targets. Here, we test the performance of the Rosetta energy function for identifying compounds that inhibit protein interactions, when these active compounds have been hidden amongst pools of "decoys." Through this virtual screening benchmark, we gauge the effect of two recent enhancements to the functional form of the Rosetta energy function: the new "Talaris" update and the "pwSHO" solvation model. Finally, we conclude by developing and validating a new weight set that maximizes Rosetta's ability to pick out the active compounds in this test set. Looking collectively over the course of these enhancements, we find a marked improvement in Rosetta's ability to identify small-molecule inhibitors of protein-protein interactions.
SN 1932-6203
PD OCT 20
PY 2015
VL 10
IS 10
AR e0140359
DI 10.1371/journal.pone.0140359
UT WOS:000363028100028
PM 26484863
ER

PT J
AU Brender, JR
Zhang, Y
AF Brender, Jeffrey R.
Zhang, Yang
TI Predicting the Effect of Mutations on Protein-Protein Binding Interactions through Structure-Based Interface Profiles
SO PLOS COMPUTATIONAL BIOLOGY
AB The formation of protein-protein complexes is essential for proteins to perform their physiological functions in the cell. Mutations that prevent the proper formation of the correct complexes can have serious consequences for the associated cellular processes. Since experimental determination of protein-protein binding affinity remains difficult when performed on a large scale, computational methods for predicting the consequences of mutations on binding affinity are highly desirable. We show that a scoring function based on interface structure profiles collected from analogous protein-protein interactions in the PDB is a powerful predictor of protein binding affinity changes upon mutation. As a standalone feature, the differences between the interface profile score of the mutant and wild-type proteins has an accuracy equivalent to the best all-atom potentials, despite being two orders of

magnitude faster once the profile has been constructed. Due to its unique sensitivity in collecting the evolutionary profiles of analogous binding interactions and the high speed of calculation, the interface profile score has additional advantages as a complementary feature to combine with physics-based potentials for improving the accuracy of composite scoring approaches. By incorporating the sequence-derived and residue-level coarse-grained potentials with the interface structure profile score, a composite model was constructed through the random forest training, which generates a Pearson correlation coefficient >0.8 between the predicted and observed binding free-energy changes upon mutation. This accuracy is comparable to, or outperforms in most cases, the current best methods, but does not require high-resolution full-atomic models of the mutant structures. The binding interface profiling approach should find useful application in human-disease mutation recognition and protein interface design studies.
SN 1553-734X
EI 1553-7358
PD OCT
PY 2015
VL 11
IS 10
AR e1004494
DI 10.1371/journal.pcbi.1004494
UT WOS:000364399700045
PM 26506533
ER

PT J
AU Cooper, GM
AF Cooper, Gregory M.
TI Parlez-vous VUS?
SO GENOME RESEARCH
AB Human genome sequencing is routine and will soon be a staple in research and clinical genetics. However, the promise of sequencing is often just that, with genome data routinely failing to reveal useful insights about disease in general or a person's health in particular. Nowhere is this chasm between promise and progress more evident than in the designation, "variant of uncertain significance" (VUS). Although it serves an important role, careful consideration of VUS reveals it to be a nebulous description of genomic information and its relationship to disease, symptomatic of our inability to make even crude quantitative assertions about the disease risks conferred by many genetic variants. In this perspective, I discuss the challenge of "variant interpretation" and the value of comparative and functional genomic information in meeting that challenge. Although already essential, genomic annotations will become even more important as our analytical focus widens beyond coding exons. Combined with more genotype and phenotype data, they will help facilitate more quantitative and insightful assessments of the contributions of genetic variants to disease.
SN 1088-9051
EI 1549-5469
PD OCT
PY 2015
VL 25
IS 10
BP 1423
EP 1426
DI 10.1101/gr.190116.115
UT WOS:000362157400004
PM 26430151
ER

PT J
AU Koenig, P
  Lee, CV
  Sanowar, S
  Wu, P
  Stinson, J
  Harris, SF
  Fuh, G
AF Koenig, Patrick
  Lee, Chingwei V.
  Sanowar, Sarah
  Wu, Ping
  Stinson, Jeremy
  Harris, Seth F.
  Fuh, Germaine
TI Deep Sequencing-guided Design of a High Affinity Dual Specificity Antibody to Target Two Angiogenic Factors in Neovascular Age-related Macular Degeneration
SO JOURNAL OF BIOLOGICAL CHEMISTRY
AB The development of dual targeting antibodies promises therapies with improved efficacy over mono-specific antibodies. Here, we engineered a Two-in-One VEGF/angiopoietin 2 antibody with dual action Fab (DAF) as a potential therapeutic for neovascular age-related macular degeneration. Crystal structures of the VEGF/angiopoietin 2 DAF in complex with its two antigens showed highly overlapping binding sites. To achieve sufficient affinity of the

DAF to block both angiogenic factors, we turned to deep mutational scanning in the complementarity determining regions (CDRs). By mutating all three CDRs of each antibody chain simultaneously, we were able not only to identify affinity improving single mutations but also mutation pairs from different CDRs that synergistically improve both binding functions. Furthermore, insights into the cooperativity between mutations allowed us to identify fold-stabilizing mutations in the CDRs. The data obtained from deep mutational scanning reveal that the majority of the 52 CDR residues are utilized differently for the two antigen binding function and permit, for the first time, the engineering of several DAF variants with subnanomolar affinity against two structurally unrelated antigens. The improved variants show similar blocking activity of receptor binding as the high affinity mono-specific antibodies against these two proteins, demonstrating the feasibility of generating a dual specificity binding surface with comparable properties to individual high affinity mono-specific antibodies.
SN 0021-9258
EI 1083-351X
PD SEP 4
PY 2015
VL 290
IS 36
BP 21773
EP 21786
DI 10.1074/jbc.M115.662783
UT WOS:000360968500001
PM 26088137
ER

PT J
AU Shin, H
  Cho, BK
AF Shin, HyeonSeok
  Cho, Byung-Kwan
TI Rational Protein Engineering Guided by Deep Mutational Scanning
SO INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES
AB Sequence-function relationship in a protein is commonly determined by the three-dimensional protein structure followed by various biochemical experiments. However, with the explosive increase in the number of genome sequences, facilitated by recent advances in sequencing technology, the gap between protein sequences available and three-dimensional structures is rapidly widening. A recently developed method termed deep mutational scanning explores the functional phenotype of thousands of mutants via massive sequencing. Coupled with a highly efficient screening system, this approach assesses the phenotypic changes made by the substitution of each amino acid sequence that constitutes a protein. Such an informational resource provides the functional role of each amino acid sequence, thereby providing sufficient rationale for selecting target residues for protein engineering. Here, we discuss the current applications of deep mutational scanning and consider experimental design.
RI Cho, Byung-Kwan/C-1830-2011
SN 1422-0067
PD SEP
PY 2015
VL 16
IS 9
BP 23094
EP 23110
DI 10.3390/ijms160923094
UT WOS:000364541000164
PM 26404267
ER

PT J
AU Cho, N
  Hwang, B
  Yoon, JK
  Park, S
  Lee, J
  Seo, HN
  Lee, J
  Huh, S
  Chung, J
  Bang, D
AF Cho, Namjin
  Hwang, Byungjin
  Yoon, Jung-ki
  Park, Sangun
  Lee, Joongoo
  Seo, Han Na
  Lee, Jeewon
  Huh, Sunghoon

Chung, Jinsoo
Bang, Duhee
TI y De novo assembly and next-generation sequencing to analyse full-length gene variants from codon-barcoded libraries
SO NATURE COMMUNICATIONS
AB Interpreting epistatic interactions is crucial for understanding evolutionary dynamics of complex genetic systems and unveiling structure and function of genetic pathways. Although high resolution mapping of en masse variant libraries renders molecular biologists to address genotype-phenotype relationships, long-read sequencing technology remains indispensable to assess functional relationship between mutations that lie far apart. Here, we introduce JigsawSeq for multiplexed sequence identification of pooled gene variant libraries by combining a codon-based molecular barcoding strategy and de novo assembly of short-read data. We first validate JigsawSeq on small sub-pools and observed high precision and recall at various experimental settings. With extensive simulations, we then apply JigsawSeq to large-scale gene variant libraries to show that our method can be reliably scaled using next-generation sequencing. JigsawSeq may serve as a rapid screening tool for functional genomics and offer the opportunity to explore evolutionary trajectories of protein variants.
SN 2041-1723
PD SEP
PY 2015
VL 6
AR 8351
DI 10.1038/ncomms9351
UT WOS:000363022000006
PM 26387459
ER

PT J
AU Smyth, RP
Despons, L
Gong, HL
Bernacchi, S
Hijnen, M
Mak, J
Jossinet, F
Li, WX
Paillart, JC
von Kleist, M
Marquet, R
AF Smyth, Redmond P.
Despons, Laurence
Gong Huili
Bernacchi, Serena
Hijnen, Marcel
Mak, Johnson
Jossinet, Fabrice
Li Weixi
Paillart, Jean-Christophe
von Kleist, Max
Marquet, Roland
TI Mutational interference mapping experiment (MIME) for studying RNA structure and function
SO NATURE METHODS
AB RNA regulates many biological processes; however, identifying functional RNA sequences and structures is complex and time-consuming. We introduce a method, mutational interference mapping experiment (MIME), to identify, at single-nucleotide resolution, the primary sequence and secondary structures of an RNA molecule that are crucial for its function. MIME is based on random mutagenesis of the RNA target followed by functional selection and next-generation sequencing. Our analytical approach allows the recovery of quantitative binding parameters and permits the identification of base-pairing partners directly from the sequencing data. We used this method to map the binding site of the human immunodeficiency virus-1 (HIHIV-1) Pr55(Gag) protein on the viral genomic RNA in vitro, and showed that, by analyzing permitted base-pairing patterns, we could model RNA structure motifs that are crucial for protein binding.
RI mak, johnson/H-4605-2014; von Kleist, Max/K-3023-2016;
OI mak, johnson/0000-0002-5229-5707; von Kleist, Max/0000-0001-6587-6394;
Marquet, Roland/0000-0002-4209-3976; Paillart,
jean-christophe/0000-0003-1647-8917; Smyth, Redmond/0000-0002-1580-0671
SN 1548-7091
EI 1548-7105
PD SEP
PY 2015
VL 12
IS 9
BP 866

EP +
DI 10.1038/NMETH.3490
UT WOS:000360586700033
PM 26237229
ER

PT J
AU Christiansen, A
Kringelum, JV
Hansen, CS
Bogh, KL
Sullivan, E
Patel, J
Rigby, NM
Eiwegger, T
Szepfalusi, Z
de Masi, F
Nielsen, M
Lund, O
Dufva, M
AF Christiansen, Anders
Kringelum, Jens V.
Hansen, Christian S.
Bogh, Katrine L.
Sullivan, Eric
Patel, Jigar
Rigby, Neil M.
Eiwegger, Thomas
Szepfalusi, Zsolt
de Masi, Federico
Nielsen, Morten
Lund, Ole
Dufva, Martin
TI High-throughput sequencing enhanced phage display enables the identification of patient-specific epitope motifs in serum
SO SCIENTIFIC REPORTS
AB Phage display is a prominent screening technique with a multitude of applications including therapeutic antibody development and mapping of antigen epitopes. In this study, phages were selected based on their interaction with patient serum and exhaustively characterised by high-throughput sequencing. A bioinformatics approach was developed in order to identify peptide motifs of interest based on clustering and contrasting to control samples. Comparison of patient and control samples confirmed a major issue in phage display, namely the selection of unspecific peptides. The potential of the bioinformatic approach was demonstrated by identifying epitopes of a prominent peanut allergen, Ara h 1, in sera from patients with severe peanut allergy. The identified epitopes were confirmed by high-density peptide micro-arrays. The present study demonstrates that high-throughput sequencing can empower phage display by (i) enabling the analysis of complex biological samples, (ii) circumventing the traditional laborious picking and functional testing of individual phage clones and (iii) reducing the number of selection rounds.
RI Dufva, Martin /D-1873-2012
OI Dufva, Martin /0000-0001-5449-0189
SN 2045-2322
PD AUG 6
PY 2015
VL 5
AR 12913
DI 10.1038/srep12913
UT WOS:000359126700001
PM 26246327
ER

PT J
AU Magliery, TJ
AF Magliery, Thomas J.
TI Protein stability: computation, sequence statistics, and new experimental methods
SO CURRENT OPINION IN STRUCTURAL BIOLOGY
AB Calculating protein stability and predicting stabilizing mutations remain exceedingly difficult tasks, largely due to the inadequacy of potential functions, the difficulty of modeling entropy and the unfolded state, and challenges of sampling, particularly of backbone conformations. Yet, computational design has produced some remarkably stable proteins in recent years, apparently owing to near ideality in structure and sequence features. With caveats, computational prediction of stability can be used to guide mutation, and mutations derived from consensus sequence analysis, especially improved by recent co-variation filters, are very likely to stabilize without sacrificing function. The combination of computational and statistical

approaches with library approaches, including new technologies such as deep sequencing and high throughput stability measurements, point to a very exciting near term future for stability engineering, even with difficult computational issues remaining.
SN 0959-440X
EI 1879-033X
PD AUG
PY 2015
VL 33
BP 161
EP 168
DI 10.1016/j.sbi.2015.09.002
UT WOS:000365362400018
PM 26497286
ER

PT J
AU Rockah-Shmuel, L
Toth-Petroczy, A
Tawfik, DS
AF Rockah-Shmuel, Liat
Toth-Petroczy, Agnes
Tawfik, Dan S.
TI Systematic Mapping of Protein Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly Neutral Mutations
SO PLOS COMPUTATIONAL BIOLOGY
AB Systematic mappings of the effects of protein mutations are becoming increasingly popular. Unexpectedly, these experiments often find that proteins are tolerant to most amino acid substitutions, including substitutions in positions that are highly conserved in nature. To obtain a more realistic distribution of the effects of protein mutations, we applied a laboratory drift comprising 17 rounds of random mutagenesis and selection of M.HaeIII, a DNA methyltransferase. During this drift, multiple mutations gradually accumulated. Deep sequencing of the drifted gene ensembles allowed determination of the relative effects of all possible single nucleotide mutations. Despite being averaged across many different genetic backgrounds, about 67% of all nonsynonymous, missense mutations were evidently deleterious, and an additional 16% were likely to be deleterious. In the early generations, the frequency of most deleterious mutations remained high. However, by the 17th generation, their frequency was consistently reduced, and those remaining were accepted alongside compensatory mutations. The tolerance to mutations measured in this laboratory drift correlated with sequence exchanges seen in M. HaeIII's natural orthologs. The biophysical constraints dictating purging in nature and in this laboratory drift also seemed to overlap. Our experiment therefore provides an improved method for measuring the effects of protein mutations that more closely replicates the natural evolutionary forces, and thereby a more realistic view of the mutational space of proteins.
SN 1553-734X
EI 1553-7358
PD AUG
PY 2015
VL 11
IS 8
AR e1004421
DI 10.1371/journal.pcbi.1004421
UT WOS:000360824500032
PM 26274323
ER

PT J
AU Kretz, CA
Dai, MH
Soylemez, O
Yee, A
Desch, KC
Siemieniak, D
Tomberg, K
Kondrashov, FA
Meng, F
Ginsburg, D
AF Kretz, Colin A.
Dai, Manhong
Soylemez, Onuralp
Yee, Andrew
Desch, Karl C.
Siemieniak, David
Tomberg, Kaert
Kondrashov, Fyodor A.
Meng, Fan

Ginsburg, David
TI Massively parallel enzyme kinetics reveals the substrate recognition landscape of the metalloprotease ADAMTS13
SO PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA
AB Proteases play important roles in many biologic processes and are key mediators of cancer, inflammation, and thrombosis. However, comprehensive and quantitative techniques to define the substrate specificity profile of proteases are lacking. The metalloprotease ADAMTS13 regulates blood coagulation by cleaving vonWillebrand factor (VWF), reducing its procoagulant activity. A mutagenized substrate phage display library based on a 73-amino acid fragment of VWF was constructed, and the ADAMTS13-dependent change in library complexity was evaluated over reaction time points, using high-throughput sequencing. Reaction rate constants (k(cat)/K-M) were calculated for nearly every possible single amino acid substitution within this fragment. This massively parallel enzyme kinetics analysis detailed the specificity of ADAMTS13 and demonstrated the critical importance of the P1-P1' substrate residues while defining exosite binding domains. These data provided empirical evidence for the propensity for epistasis within VWF and showed strong correlation to conservation across orthologs, highlighting evolutionary selective pressures for VWF.
OI Soylemez, Onuralp/0000-0001-8308-6855
SN 0027-8424
PD JUL 28
PY 2015
VL 112
IS 30
BP 9328
EP 9333
DI 10.1073/pnas.1511328112
UT WOS:000358656500061
PM 26170332
ER

PT J
AU Anderson, DW
McKeown, AN
Thornton, JW
AF Anderson, Dave W.
McKeown, Alesia N.
Thornton, Joseph W.
TI Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites
SO ELIFE
AB Complexes of specifically interacting molecules, such as transcription factor proteins (TFs) and the DNA response elements (REs) they recognize, control most biological processes, but little is known concerning the functional and evolutionary effects of epistatic interactions across molecular interfaces. We experimentally characterized all combinations of genotypes in the joint protein-DNA sequence space defined by an historical transition in TF-RE specificity that occurred some 500 million years ago in the DNA-binding domain of an ancient steroid hormone receptor. We found that rampant epistasis within and between the two molecules was essential to specific TF-RE recognition and to the evolution of a novel TF-RE complex with unique derived specificity. Permissive and restrictive epistatic mutations across the TF-RE interface opened and closed potential evolutionary paths accessible by the other, making the evolution of each molecule contingent on its partner's history and allowing a molecular complex with novel specificity to evolve.
DOI: 10.7554/eLife.07864.001
SN 2050-084X
PD JUN 15
PY 2015
VL 4
AR e07864
DI 10.7554/eLife.07864
UT WOS:000373442000001
PM 26076233
ER

PT J
AU Romero, PA
Tran, TM
Abate, AR
AF Romero, Philip A.
Tran, Tuan M.
Abate, Adam R.
TI Dissecting enzyme function with microfluidic-based deep mutational scanning
SO PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA
AB Natural enzymes are incredibly proficient catalysts, but engineering them to have new or improved functions is challenging due to the complexity of how an enzyme's sequence relates to its biochemical properties. Here, we present an ultrahigh-throughput method for mapping enzyme sequence-

function relationships that combines droplet microfluidic screening with next-generation DNA sequencing. We apply our method to map the activity of millions of glycosidase sequence variants. Microfluidic-based deep mutational scanning provides a comprehensive and unbiased view of the enzyme function landscape. The mapping displays expected patterns of mutational tolerance and a strong correspondence to sequence variation within the enzyme family, but also reveals previously unreported sites that are crucial for glycosidase function. We modified the screening protocol to include a high-temperature incubation step, and the resulting thermotolerance landscape allowed the discovery of mutations that enhance enzyme thermostability. Droplet microfluidics provides a general platform for enzyme screening that, when combined with DNA-sequencing technologies, enables high-throughput mapping of enzyme sequence space.
SN 0027-8424
PD JUN 9
PY 2015
VL 112
IS 23
BP 7159
EP 7164
DI 10.1073/pnas.1422285112
UT WOS:000355823200034
PM 26040002
ER

PT J
AU Hu, DM
Hu, SY
Wan, W
Xu, M
Du, RK
Zhao, W
Gao, XL
Liu, J
Liu, HY
Hong, J
AF Hu, Dongmei
Hu, Siyi
Wan, Wen
Xu, Man
Du, Ruikai
Zhao, Wei
Gao, Xiaolian
Liu, Jing
Liu, Haiyan
Hong, Jiong
TI Effective Optimization of Antibody Affinity by Phage Display Integrated with High-Throughput DNA Synthesis and Sequencing Technologies
SO PLOS ONE
AB Phage display technology has been widely used for antibody affinity maturation for decades. The limited library sequence diversity together with excessive redundancy and labour-consuming procedure for candidate identification are two major obstacles to widespread adoption of this technology. We hereby describe a novel library generation and screening approach to address the problems. The approach started with the targeted diversification of multiple complementarity determining regions (CDRs) of a humanized anti-ErbB2 antibody, HuA21, with a small perturbation mutagenesis strategy. A combination of three degenerate codons, NWG, NWC, and NSG, were chosen for amino acid saturation mutagenesis without introducing cysteine and stop residues. In total, 7,749 degenerate oligonucleotides were synthesized on two microchips and released to construct five single-chain antibody fragment (scFv) gene libraries with 4 x 106 DNA sequences. Deep sequencing of the unselected and selected phage libraries using the Illumina platform allowed for an in-depth evaluation of the enrichment landscapes in CDR sequences and amino acid substitutions. Potent candidates were identified according to their high frequencies using NGS analysis, by-passing the need for the primary screening of target-binding clones. Furthermore, a subsequent library by recombination of the 10 most abundant variants from four CDRs was constructed and screened, and a mutant with 158-fold increased affinity (K-d = 25.5 pM) was obtained. These results suggest the potential application of the developed methodology for optimizing the binding properties of other antibodies and biomolecules.
RI Liu, Haiyan/O-9637-2014; HONG, Jiong/N-1996-2013
OI HONG, Jiong/0000-0002-4592-7083
SN 1932-6203
PD JUN 5
PY 2015
VL 10
IS 6
AR UNSP e0129125
DI 10.1371/journal.pone.0129125
UT WOS:000355652200127
PM 26046845
ER

PT J
AU Reich, L
Dutta, S
Keating, AE
AF Reich, Lothar Luther
Dutta, Sanjib
Keating, Amy E.
TI SORTCERY-A High-Throughput Method to Affinity Rank Peptide Ligands
SO JOURNAL OF MOLECULAR BIOLOGY
AB Uncovering the relationships between peptide and protein sequences and binding properties is critical for successfully predicting, re-designing and inhibiting protein-protein interactions. Systematically collected data that link protein sequence to binding are valuable for elucidating determinants of protein interaction but are rare in the literature because such data are experimentally difficult to generate. Here we describe SORTCERY, a high-throughput method that we have used to rank hundreds of yeast-displayed peptides according to their affinities for a target interaction partner. The procedure involves fluorescence-activated cell sorting of a library, deep sequencing of sorted pools and downstream computational analysis. We have developed theoretical models and statistical tools that assist in planning these stages. We demonstrate SORTCERY's utility by ranking 1026 BH3 (Bcl-2 homology 3) peptides with respect to their affinities for the anti-apoptotic protein Bcl-x(L). Our results are in striking agreement with measured affinities for 19 individual peptides with dissociation constants ranging from 0.1 to 60 nM. High-resolution ranking can be used to improve our understanding of sequence-function relationships and to support the development of computational models for predicting and designing novel interactions. (C) 2014 Elsevier Ltd. All rights reserved.
SN 0022-2836
EI 1089-8638
PD JUN 5
PY 2015
VL 427
IS 11
SI SI
BP 2135
EP 2150
DI 10.1016/j.jmb.2014.09.025
UT WOS:000355028100010
PM 25311858
ER

PT J
AU Starita, LM
Young, DL
Islam, M
Kitzman, JO
Gullingsrud, J
Hause, RJ
Fowler, DM
Parvin, JD
Shendure, J
Fields, S
AF Starita, Lea M.
Young, David L.
Islam, Muhtadi
Kitzman, Jacob O.
Gullingsrud, Justin
Hause, Ronald J.
Fowler, Douglas M.
Parvin, Jeffrey D.
Shendure, Jay
Fields, Stanley
TI Massively Parallel Functional Analysis of BRCA1 RING Domain Variants
SO GENETICS
AB Interpreting variants of uncertain significance (VUS) is a central challenge in medical genetics. One approach is to experimentally measure the functional consequences of VUS, but to date this approach has been post hoc and low throughput. Here we use massively parallel assays to measure the effects of nearly 2000 missense substitutions in the RING domain of BRCA1 on its E3 ubiquitin ligase activity and its binding to the BARD1 RING domain. From the resulting scores, we generate a model to predict the capacities of full-length BRCA1 variants to support homology-directed DNA repair, the essential role of BRCA1 in tumor suppression, and show that it outperforms widely used biological-effect prediction algorithms. We envision that massively parallel functional assays may facilitate the prospective interpretation of variants observed in clinical sequencing.
RI Parvin, Jeffrey/C-8955-2009
SN 0016-6731
EI 1943-2631
PD JUN

PY 2015
VL 200
IS 2
BP 413
EP +
DI 10.1534/genetics.115.175802
UT WOS:000356509100003
PM 25823446
ER

PT J
AU Bloom, JD
AF Bloom, Jesse D.
TI Software for the analysis and visualization of deep mutational scanning data
SO BMC BIOINFORMATICS
AB Background: Deep mutational scanning is a technique to estimate the impacts of mutations on a gene by using deep sequencing to count mutations in a library of variants before and after imposing a functional selection. The impacts of mutations must be inferred from changes in their counts after selection.
Results: I describe a software package, dms_tools, to infer the impacts of mutations from deep mutational scanning data using a likelihood-based treatment of the mutation counts. I show that dms_tools yields more accurate inferences on simulated data than simply calculating ratios of counts pre- and post-selection. Using dms_tools, one can infer the preference of each site for each amino acid given a single selection pressure, or assess the extent to which these preferences change under different selection pressures. The preferences and their changes can be intuitively visualized with sequence-logo-style plots created using an extension to weblogo.
Conclusions: dms_tools implements a statistically principled approach for the analysis and subsequent visualization of deep mutational scanning data.
RI Bloom, Jesse/C-6837-2013
OI Bloom, Jesse/0000-0003-1267-3408
SN 1471-2105
PD MAY 20
PY 2015
VL 16
AR 168
DI 10.1186/s12859-015-0590-4
UT WOS:000354972000001
PM 25990960
ER

PT J
AU Van Blarcom, T
Rossi, A
Foletti, D
Sundar, P
Pitts, S
Bee, C
Witt, JM
Melton, Z
Hasa-Moreno, A
Shaughnessy, L
Telman, D
Zhao, L
Cheung, WL
Berka, J
Zhai, WW
Strop, P
Chaparro-Riggers, J
Shelton, DL
Pons, J
Rajpal, A
AF Van Blarcom, Thomas
Rossi, Andrea
Foletti, Davide
Sundar, Purnima
Pitts, Steven
Bee, Christine
Witt, Jody Melton
Melton, Zea
Hasa-Moreno, Adela

Shaughnessy, Lee
Telman, Dilduz
Zhao, Lora
Cheung, Wai Ling
Berka, Jan
Zhai, Wenwu
Strop, Pavel
Chaparro-Riggers, Javier
Shelton, David L.
Pons, Jaume
Rajpal, Arvind
TI Precise and Efficient Antibody Epitope Determination through Library Design, Yeast Display and Next-Generation Sequencing
SO JOURNAL OF MOLECULAR BIOLOGY
AB The ability of antibodies to bind an antigen with a high degree of affinity and specificity has led them to become the largest and fastest growing class of therapeutic proteins. Clearly identifying the epitope at which they bind their cognate antigen provides insight into their mechanism of action and helps differentiate antibodies that bind the same antigen. Here, we describe a method to precisely and efficiently map the epitopes of a panel of antibodies in parallel over the course of several weeks. This method relies on the combination of rational library design, quantitative yeast surface display and next-generation DNA sequencing and was demonstrated by mapping the epitopes of several antibodies that neutralize alpha toxin from Staphylococcus aureus. The accuracy of this method was confirmed by comparing the results to the co-crystal structure of one antibody and alpha toxin and was further refined by the inclusion of a lower-affinity variant of the antibody. In addition, this method produced quantitative insight into the epitope residues most critical for the antibody antigen interaction and enabled the relative affinities of each antibody toward alpha toxin variants to be estimated. This affinity estimate serves as a predictor of neutralizing antibody potency and was used to anticipate the ability of each antibody to effectively bind and neutralize naturally occurring alpha toxin variants secreted by strains of S. aureus, including clinically relevant strains. Ultimately this type information can be used to help select the best clinical candidate among a set of antibodies against a given antigen. (C) 2014 Elsevier Ltd. All rights reserved.
SN 0022-2836
EI 1089-8638
PD MAR 27
PY 2015
VL 427
IS 6
BP 1513
EP 1534
DI 10.1016/j.jmb.2014.09.020
PN B
UT WOS:000351798700022
PM 25284753
ER

PT J
AU Kowalsky, CA
Klesmith, JR
Stapleton, JA
Kelly, V
Reichkitzer, N
Whitehead, TA
AF Kowalsky, Caitlin A.
Klesmith, Justin R.
Stapleton, James A.
Kelly, Vince
Reichkitzer, Nolan
Whitehead, Timothy A.
TI High-Resolution Sequence-Function Mapping of Full-Length Proteins
SO PLOS ONE
AB Comprehensive sequence-function mapping involves detailing the fitness contribution of every possible single mutation to a gene by comparing the abundance of each library variant before and after selection for the phenotype of interest. Deep sequencing of library DNA allows frequency reconstruction for tens of thousands of variants in a single experiment, yet short read lengths of current sequencers makes it challenging to probe genes encoding full-length proteins. Here we extend the scope of sequence-function maps to entire protein sequences with a modular, universal sequence tiling method. We demonstrate the approach with both growth-based selections and FACS screening, offer parameters and best practices that simplify design of experiments, and present analytical solutions to normalize data across independent selections. Using this protocol, sequence-function maps covering full sequences can be obtained in four to six weeks. Best practices introduced in this manuscript are fully compatible with, and complementary to, other recently published sequence-function mapping protocols.
OI Klesmith, Justin/0000-0003-2908-9355
SN 1932-6203
PD MAR 19
PY 2015
VL 10

IS 3
AR e0118193
DI 10.1371/journal.pone.0118193
UT WOS:000351425400025
PM 25790064
ER

PT J
AU Currin, A
Swainston, N
Day, PJ
Kell, DB
AF Currin, Andrew
Swainston, Neil
Day, Philip J.
Kell, Douglas B.
TI Synthetic biology for the directed evolution of protein biocatalysts:
navigating sequence space intelligently
SO CHEMICAL SOCIETY REVIEWS
AB The amino acid sequence of a protein affects both its structure and its function. Thus, the ability to modify the sequence, and hence the structure and activity, of individual proteins in a systematic way, opens up many opportunities, both scientifically and (as we focus on here) for exploitation in biocatalysis. Modern methods of synthetic biology, whereby increasingly large sequences of DNA can be synthesised de novo, allow an unprecedented ability to engineer proteins with novel functions. However, the number of possible proteins is far too large to test individually, so we need means for navigating the `search space' of possible protein sequences efficiently and reliably in order to find desirable activities and other properties. Enzymologists distinguish binding (K-d) and catalytic (k(cat)) steps. In a similar way, judicious strategies have blended design (for binding, specificity and active site modelling) with the more empirical methods of classical directed evolution (DE) for improving kcat (where natural evolution rarely seeks the highest values), especially with regard to residues distant from the active site and where the functional linkages underpinning enzyme dynamics are both unknown and hard to predict. Epistasis (where the `best' amino acid at one site depends on that or those at others) is a notable feature of directed evolution. The aim of this review is to highlight some of the approaches that are being developed to allow us to use directed evolution to improve enzyme properties, often dramatically. We note that directed evolution differs in a number of ways from natural evolution, including in particular the available mechanisms and the likely selection pressures. Thus, we stress the opportunities afforded by techniques that enable one to map sequence to (structure and) activity in silico, as an effective means of modelling and exploring protein landscapes. Because known landscapes may be assessed and reasoned about as a whole, simultaneously, this offers opportunities for protein improvement not readily available to natural evolution on rapid timescales. Intelligent landscape navigation, informed by sequence-activity relationships and coupled to the emerging methods of synthetic biology, offers scope for the development of novel biocatalysts that are both highly active and robust.
SN 0306-0012
EI 1460-4744
PD MAR 7
PY 2015
VL 44
IS 5
BP 1172
EP 1239
DI 10.1039/c4cs00351a
UT WOS:000350568000008
PM 25503938
ER

PT J
AU Alam, KK
Chang, JL
Burke, DH
AF Alam, Khalid K.
Chang, Jonathan L.
Burke, Donald H.
TI FASTAptamer: A Bioinformatic Toolkit for High-throughput Sequence
Analysis of Combinatorial Selections
SO MOLECULAR THERAPY-NUCLEIC ACIDS
AB High-throughput sequence (HTS) analysis of combinatorial selection populations accelerates lead discovery and optimization and offers dynamic insight into selection processes. An underlying principle is that selection enriches high-fitness sequences as a fraction of the population, whereas low-fitness sequences are depleted. HTS analysis readily provides the requisite numerical information by tracking the evolutionary trajectory of individual sequences in response to selection pressures. Unlike genomic data, for which a number of software solutions exist, user-friendly tools are not readily available for the combinatorial selections field, leading many users to create custom software. FASTAptamer was designed to address the sequence-level analysis needs of the field. The open source FASTAptamer toolkit counts, normalizes and ranks read counts in a FASTQ file, compares populations for sequence distribution, generates clusters of sequence families, calculates fold-enrichment of sequences throughout the course of a selection and searches for degenerate sequence motifs. While originally designed for aptamer selections, FASTAptamer can be applied to any selection strategy that can utilize next-generation DNA sequencing, such as ribozyme or deoxyribozyme selections, in vivo mutagenesis and various surface display technologies

(peptide, antibody fragment, mRNA, etc.).
SN 2162-2531
PD MAR
PY 2015
VL 4
AR e230
DI 10.1038/mtna.2015.4
UT WOS:000358344200001
PM 25734917
ER

PT J
AU Stiffler, MA
Hekstra, DR
Ranganathan, R
AF Stiffler, Michael A.
Hekstra, Doeke R.
Ranganathan, Rama
TI Evolvability as a Function of Purifying Selection in TEM-1 beta-Lactamase
SO CELL
AB Evolvability-the capacity to generate beneficial heritable variation-is a central property of biological systems. However, its origins and modulation by environmental factors have not been examined systematically. Here, we analyze the fitness effects of all single mutations in TEM-1 beta-lactamase (4,997 variants) under selection for the wild-type function (ampicillin resistance) and for a new function (cefotaxime resistance). Tolerance to mutation in this enzyme is bimodal and dependent on the strength of purifying selection in vivo, a result that derives from a steep non-linear ampicillin-dependent relationship between biochemical activity and fitness. Interestingly, cefotaxime resistance emerges from mutations that are neutral at low levels of ampicillin but deleterious at high levels; thus the capacity to evolve new function also depends on the strength of selection. The key property controlling evolvability is an excess of enzymatic activity relative to the strength of selection, suggesting that fluctuating environments might select for high-activity enzymes.
SN 0092-8674
EI 1097-4172
PD FEB 26
PY 2015
VL 160
IS 5
BP 882
EP 892
DI 10.1016/j.cell.2015.01.035
UT WOS:000351116100011
PM 25723163
ER

PT J
AU Abriata, LA
Palzkill, T
Dal Peraro, M
AF Abriata, Luciano A.
Palzkill, Timothy
Dal Peraro, Matteo
TI How Structural and Physicochemical Determinants Shape Sequence Constraints in a Functional Enzyme
SO PLOS ONE
AB The need for interfacing structural biology and biophysics to molecular evolution is being increasingly recognized. One part of the big problem is to understand how physics and chemistry shape the sequence space available to functional proteins, while satisfying the needs of biology. Here we present a quantitative, structure-based analysis of a high-resolution map describing the tolerance to all substitutions in all positions of a functional enzyme, namely a TEM lactamase previously studied through deep sequencing of mutants growing in competition experiments with selection against ampicillin. Substitutions are rarely observed within 7 angstrom of the active site, a stringency that is relaxed slowly and extends up to 15-20 20 angstrom, with buried residues being especially sensitive. Substitution patterns in over one third of the residues can be quantitatively modeled by monotonic dependencies on amino acid descriptors and predictions of changes in folding stability. Amino acid volume and steric hindrance shape constraints on the protein core; hydrophobicity and solubility shape constraints on hydrophobic clusters underneath the surface, and on salt bridges and polar networks at the protein surface together with charge and hydrogen bonding capacity. Amino acid solubility, flexibility and conformational descriptors also provide additional constraints at many locations. These findings provide fundamental insights into the chemistry underlying protein evolution and design, by quantitating links between sequence and different protein traits, illuminating subtle and unexpected sequence-trait relationships and pinpointing what traits are sacrificed upon gain-of-function mutation.
SN 1932-6203
PD FEB 23
PY 2015

VL 10
IS 2
AR e0118684
DI 10.1371/journal.pone.0118684
UT WOS:000350662100201
PM 25706742
ER

PT J
AU Podgornaia, AI
Laub, MT
AF Podgornaia, Anna I.
Laub, Michael T.
TI Pervasive degeneracy and epistasis in a protein-protein interface
SO SCIENCE
AB Mapping protein sequence space is a difficult problem that necessitates the analysis of 20(N) combinations for sequences of length N. We systematically mapped the sequence space of four key residues in the Escherichia coli protein kinase PhoQ that drive recognition of its substrate PhoP. We generated a library containing all 160,000 variants of PhoQ at these positions and used a two-step selection coupled to next-generation sequencing to identify 1659 functional variants. Our results reveal extensive degeneracy in the PhoQ-PhoP interface and epistasis, with the effect of individual substitutions often highly dependent on context. Together, epistasis and the genetic code create a pattern of connectivity of functional variants in sequence space that likely constrains PhoQ evolution. Consequently, the diversity of PhoQ orthologs is substantially lower than that of functional PhoQ variants.
SN 0036-8075
EI 1095-9203
PD FEB 6
PY 2015
VL 347
IS 6222
BP 673
EP 677
DI 10.1126/science.1257360
UT WOS:000349145200047
PM 25657251
ER

PT J
AU Melamed, D
Young, DL
Miller, CR
Fields, S
AF Melamed, Daniel
Young, David L.
Miller, Christina R.
Fields, Stanley
TI Combining Natural Sequence Variation with High Throughput Mutational Data to Reveal Protein Interaction Sites
SO PLOS GENETICS
AB Many protein interactions are conserved among organisms despite changes in the amino acid sequences that comprise their contact sites, a property that has been used to infer the location of these sites from protein homology. In an inter-species complementation experiment, a sequence present in a homologue is substituted into a protein and tested for its ability to support function. Therefore, substitutions that inhibit function can identify interaction sites that changed over evolution. However, most of the sequence differences within a protein family remain unexplored because of the small-scale nature of these complementation approaches. Here we use existing high throughput mutational data on the in vivo function of the RRM2 domain of the Saccharomyces cerevisiae poly(A)-binding protein, Pab1, to analyze its sites of interaction. Of 197 single amino acid differences in 52 Pab1 homologues, 17 reduce the function of Pab1 when substituted into the yeast protein. The majority of these deleterious mutations interfere with the binding of the RRM2 domain to eIF4G1 and eIF4G2, isoforms of a translation initiation factor. A large-scale mutational analysis of the RRM2 domain in a two-hybrid assay for eIF4G1 binding supports these findings and identifies peripheral residues that make a smaller contribution to eIF4G1 binding. Three single amino acid substitutions in yeast Pab1 corresponding to residues from the human orthologue are deleterious and eliminate binding to the yeast eIF4G isoforms. We create a triple mutant that carries these substitutions and other humanizing substitutions that collectively support a switch in binding specificity of RRM2 from the yeast eIF4G1 to its human orthologue. Finally, we map other deleterious substitutions in Pab1 to inter-domain (RRM2-RRM1) or protein-RNA (RRM2-poly(A)) interaction sites. Thus, the combined approach of large-scale mutational data and evolutionary conservation can be used to characterize interaction sites at single amino acid resolution.
SN 1553-7390
EI 1553-7404
PD FEB
PY 2015
VL 11
IS 2

AR e1004918
DI 10.1371/journal.pgen.1004918
UT WOS:000352081800009
PM 25671604
ER

PT J
AU Doolan, KM
Colby, DW
AF Doolan, Kyle M.
Colby, David W.
TI Conformation-Dependent Epitopes Recognized by Prion Protein Antibodies Probed Using Mutational Scanning and Deep Sequencing
SO JOURNAL OF MOLECULAR BIOLOGY
AB Prion diseases are caused by a structural rearrangement of the cellular prion protein, PrPc, into a disease-associated conformation, PrPsc, which may be distinguished from one another using conformation-specific antibodies. We used mutational scanning by cell-surface display to screen 1341 PrP single point mutants for attenuated interaction with four anti-PrP antibodies, including several with conformational specificity. Single-molecule real-time gene sequencing was used to quantify enrichment of mutants, returning 26,000 high-quality full-length reads for each screened population on average. Relative enrichment of mutants correlated to the magnitude of the change in binding affinity. Mutations that diminished binding of the antibody ICSM18 represented the core of contact residues in the published crystal structure of its complex. A similarly located binding site was identified for D18, comprising discontinuous residues in helix 1 of PrP, brought into close proximity to one another only when the alpha helix is intact. The specificity of these antibodies for the normal form of PrP likely arises from loss of this conformational feature after conversion to the disease-associated form. Intriguingly, 6H4 binding was found to depend on interaction with the same residues, among others, suggesting that its ability to recognize both forms of PrP depends on a structural rearrangement of the antigen. The application of mutational scanning and deep sequencing provides residue-level resolution of positions in the protein protein interaction interface that are critical for binding, as well as a quantitative measure of the impact of mutations on binding affinity. (C) 2014 Elsevier Ltd. All rights reserved.
OI Doolan, Kyle/0000-0001-8449-4826
SN 0022-2836
EI 1089-8638
PD JAN 30
PY 2015
VL 427
IS 2
BP 328
EP 340
DI 10.1016/j.jmb.2014.10.024
UT WOS:000348888200013
PM 25451031
ER

PT J
AU Shaw, CA
Campbell, IM
AF Shaw, Chad A.
Campbell, Ian M.
TI Variant interpretation through Bayesian fusion of frequency and genomic knowledge
SO GENOME MEDICINE
AB Variant interpretation is a central challenge in genomic medicine. A recent study demonstrates the power of Bayesian statistical approaches to improve interpretation of variants in the context of specific genes and syndromes. Such Bayesian approaches combine frequency (in the form of observed genetic variation in cases and controls) with biological annotations to determine a probability of pathogenicity. These Bayesian approaches complement other efforts to catalog human variation.
SN 1756-994X
PD JAN 28
PY 2015
VL 7
AR 4
DI 10.1186/s13073-015-0129-3
UT WOS:000348868700001
PM 25632303
ER

PT S
AU Janin, J
Wodak, SJ
Lensink, MF
Velankar, S

AF Janin, Joel
Wodak, Shoshana J.
Lensink, Marc F.
Velankar, Sameer
BE Parrill, AL
Lipkowitz, KB
TI Assessing Structural Predictions of Protein-Protein Recognition: The CAPRI Experiment
SO REVIEWS IN COMPUTATIONAL CHEMISTRY, VOL 28
SE Reviews in Computational Chemistry
SN 1069-3599
BN 978-1-118-88988-6; 978-1-118-40777-6
PY 2015
VL 28
BP 137
EP 173
D2 10.1002/9781118889886
UT WOS:000372183900005
ER

PT J
AU Al-Mawsawi, LQ
Wu, NC
Olson, CA
Shi, VC
Qi, HF
Zheng, XJ
Wu, TT
Sun, R
AF Al-Mawsawi, Laith Q.
Wu, Nicholas C.
Olson, C. Anders
Shi, Vivian Cai
Qi, Hangfei
Zheng, Xiaojuan
Wu, Ting-Ting
Sun, Ren
TI High-throughput profiling of point mutations across the HIV-1 genome
SO RETROVIROLOGY
AB Background: The HIV-1 pandemic is not the result of a static pathogen but a large genetically diverse and dynamic viral population. The virus is characterized by a highly mutable genome rendering efforts to design a universal vaccine a significant challenge and drives the emergence of drug resistant variants upon antiviral pressure. Gaining a comprehensive understanding of the mutational tolerance of each HIV-1 genomic position is therefore of critical importance.
Results: Here we combine high-density mutagenesis with the power of next-generation sequencing to gauge the replication capacity and therefore mutational tolerability of single point mutations across the entire HIV-1 genome. We were able to achieve the evaluation of point mutational effects on viral replicative capacity for 5,553 individual HIV-1 nucleotide positions - representing 57% of the viral genome. Replicative capacity was assessed at 3,943 nucleotide positions for a single alternate base change, 1,459 nucleotide positions for two alternate base changes, and 151 nucleotide positions for all three possible alternate base changes. This resulted in the study of how a total of 7,314 individual point mutations impact HIV-1 replication on a single experimental platform. We further utilize the dataset for a focused structural analysis on a capsid inhibitor binding pocket.
Conclusion: The approach presented here can be applied to any pathogen that can be genetically manipulated in a laboratory setting. Furthermore, the methodology can be utilized under externally applied selection conditions, such as drug or immune pressure, to identify genetic elements that contribute to drug or host interactions, and therefore mutational routes of pathogen resistance and escape.
RI Wu, Nicholas/H-3822-2015
OI Wu, Nicholas/0000-0002-9078-6697
SN 1742-4690
PD DEC 19
PY 2014
VL 11
AR 124
DI 10.1186/s12977-014-0124-6
UT WOS:000349351600001
PM 25522661
ER

PT J
AU Thyme, SB
Song, YF

Brunette, TJ
Szeto, MD
Kusak, L
Bradley, P
Baker, D
AF Thyme, Summer B.
Song, Yifan
Brunette, T. J.
Szeto, Mindy D.
Kusak, Lara
Bradley, Philip
Baker, David
TI Massively parallel determination and modeling of endonuclease substrate specificity
SO NUCLEIC ACIDS RESEARCH
AB We describe the identification and characterization of novel homing endonucleases using genome database mining to identify putative target sites, followed by high throughput activity screening in a bacterial selection system. We characterized the substrate specificity and kinetics of these endonucleases by monitoring DNA cleavage events with deep sequencing. The endonuclease specificities revealed by these experiments can be partially recapitulated using 3D structure-based computational models. Analysis of these models together with genome sequence data provide insights into how alternative endonuclease specificities were generated during natural evolution.
RI Baker, David/K-8941-2012
OI Baker, David/0000-0001-7896-6217
SN 0305-1048
EI 1362-4962
PD DEC 16
PY 2014
VL 42
IS 22
BP 13839
EP 13852
DI 10.1093/nar/gku1096
UT WOS:000347916900040
PM 25389263
ER

PT J
AU Warszawski, S
Netzer, R
Tawfik, DS
Fleishman, SJ
AF Warszawski, Shira
Netzer, Ravit
Tawfik, Dan S.
Fleishman, Sarel J.
TI A "Fuzzy"-Logic Language for Encoding Multiple Physical Traits in Biomolecules
SO JOURNAL OF MOLECULAR BIOLOGY
AB To carry out their activities, biological macromolecules balance different physical traits, such as stability, interaction affinity, and selectivity. How such often opposing traits are encoded in a macromolecular system is critical to our understanding of evolutionary processes and ability to design new molecules with desired functions. We present a framework for constraining design simulations to balance different physical characteristics. Each trait is represented by the equilibrium fractional occupancy of the desired state relative to its alternatives, ranging from none to full occupancy, and the different traits are combined using Boolean operators to effect a "fuzzy"-logic language for encoding any combination of traits. In another paper, we presented a new combinatorial backbone design algorithm AbDesign where the fuzzy-logic framework was used to optimize protein backbones and sequences for both stability and binding affinity in antibody-design simulation. We now extend this framework and find that fuzzy-logic design simulations reproduce sequence and structure design principles seen in nature to underlie exquisite specificity on the one hand and multispecificity on the other hand. The fuzzy-logic language is broadly applicable and could help define the space of tolerated and beneficial mutations in natural biomolecular systems and design artificial molecules that encode complex characteristics. (C) 2014 MRC Laboratory of Molecular Biology. Published by Elsevier Ltd.
OI Fleishman, Sarel/0000-0003-3177-7560; Netzer, Ravit/0000-0003-3483-3927
SN 0022-2836
EI 1089-8638
PD DEC 12
PY 2014
VL 426
IS 24
BP 4125
EP 4138

DI 10.1016/j.jmb.2014.10.002
UT WOS:000347657000017
PM 25311857
ER

PT J
AU Ashworth, J
Bernard, B
Reynolds, S
Plaisier, CL
Shmulevich, I
Baliga, NS
AF Ashworth, Justin
Bernard, Brady
Reynolds, Sheila
Plaisier, Christopher L.
Shmulevich, Ilya
Baliga, Nitin S.
TI Structure-based predictions broadly link transcription factor mutations to gene expression changes in cancers
SO NUCLEIC ACIDS RESEARCH
AB Thousands of unique mutations in transcription factors (TFs) arise in cancers, and the functional and biological roles of relatively few of these have been characterized. Here, we used structure-based methods developed specifically for DNA-binding proteins to systematically predict the consequences of mutations in several TFs that are frequently mutated in cancers. The explicit consideration of protein-DNA interactions was crucial to explain the roles and prevalence of mutations in TP53 and RUNX1 in cancers, and resulted in a higher specificity of detection for known p53-regulated genes among genetic associations between TP53 genotypes and genome-wide expression in The Cancer Genome Atlas, compared to existing methods of mutation assessment. Biophysical predictions also indicated that the relative prevalence of TP53 missense mutations in cancer is proportional to their thermodynamic impacts on protein stability and DNA binding, which is consistent with the selection for the loss of p53 transcriptional function in cancers. Structure and thermodynamics-based predictions of the impacts of missense mutations that focus on specific molecular functions may be increasingly useful for the precise and large-scale inference of aberrant molecular phenotypes in cancer and other complex diseases.
OI Plaisier, Christohper/0000-0003-3273-5717
SN 0305-1048
EI 1362-4962
PD DEC 1
PY 2014
VL 42
IS 21
BP 12973
EP 12983
DI 10.1093/nar/gku1031
UT WOS:000347914600009
PM 25378323
ER

PT J
AU Raman, S
Taylor, N
Genuth, N
Fields, S
Church, GM
AF Raman, Srivatsan
Taylor, Noah
Genuth, Naomi
Fields, Stanley
Church, George M.
TI Engineering allostery
SO TRENDS IN GENETICS
AB Allosteric proteins have great potential in synthetic biology, but our limited understanding of the molecular underpinnings of allostery has hindered the development of designer molecules, including transcription factors with new DNA-binding or ligand-binding specificities that respond appropriately to inducers. Such allosteric proteins could function as novel switches in complex circuits, metabolite sensors, or as orthogonal regulators for independent, inducible control of multiple genes. Advances in DNA synthesis and next-generation sequencing technologies have enabled the assessment of millions of mutants in a single experiment, providing new opportunities to study allostery. Using the classic Lad protein as an example, we describe a genetic selection system using a bidirectional reporter to capture mutants in both allosteric states, allowing the positions most crucial for allostery to be identified. This approach is not limited to bacterial transcription factors, and could reveal new mechanistic insights and facilitate engineering of other major classes of allosteric proteins such as nuclear receptors, two-component systems, G protein-coupled receptors, and protein kinases.
SN 0168-9525
PD DEC

PY 2014
VL 30
IS 12
BP 521
EP 528
DI 10.1016/j.tig.2014.09.004
UT WOS:000347499500005
PM 25306102
ER

PT J
AU Wei, XM
Das, J
Fragoza, R
Liang, J
de Oliveira, FMB
Lee, HR
Wang, XJ
Mort, M
Stenson, PD
Cooper, DN
Lipkin, SM
Smolka, MB
Yu, HY
AF Wei, Xiaomu
Das, Jishnu
Fragoza, Robert
Liang, Jin
de Oliveira, Francisco M. Bastos
Lee, Hao Ran
Wang, Xiujuan
Mort, Matthew
Stenson, Peter D.
Cooper, David N.
Lipkin, Steven M.
Smolka, Marcus B.
Yu, Haiyuan
TI A Massively Parallel Pipeline to Clone DNA Variants and Examine Molecular Phenotypes of Human Disease Mutations
SO PLOS GENETICS
AB Understanding the functional relevance of DNA variants is essential for all exome and genome sequencing projects. However, current mutagenesis cloning protocols require Sanger sequencing, and thus are prohibitively costly and labor-intensive. We describe a massively-parallel site-directed mutagenesis approach, "Clone-seq", leveraging next-generation sequencing to rapidly and cost-effectively generate a large number of mutant alleles. Using Clone-seq, we further develop a comparative interactome-scanning pipeline integrating high-throughput GFP, yeast two-hybrid (Y2H), and mass spectrometry assays to systematically evaluate the functional impact of mutations on protein stability and interactions. We use this pipeline to show that disease mutations on protein-protein interaction interfaces are significantly more likely than those away from interfaces to disrupt corresponding interactions. We also find that mutation pairs with similar molecular phenotypes in terms of both protein stability and interactions are significantly more likely to cause the same disease than those with different molecular phenotypes, validating the in vivo biological relevance of our high-throughput GFP and Y2H assays, and indicating that both assays can be used to determine candidate disease mutations in the future. The general scheme of our experimental pipeline can be readily expanded to other types of interactome-mapping methods to comprehensively evaluate the functional relevance of all DNA variants, including those in non-coding regions.
RI Das, Jishnu/C-6924-2015; Bastos de Oliveira, Francisco/I-3540-2013;
Cooper, David N./H-4384-2011
OI Das, Jishnu/0000-0002-5747-064X; Cooper, David N./0000-0002-8943-8484
SN 1553-7390
EI 1553-7404
PD DEC
PY 2014
VL 10
IS 12
AR e1004819
DI 10.1371/journal.pgen.1004819
UT WOS:000346649900024
PM 25502805
ER

PT J

AU Olson, CA
Wu, NC
Sun, R
AF Olson, C. Anders
Wu, Nicholas C.
Sun, Ren
TI A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain
SO CURRENT BIOLOGY
AB Background: Nonadditivity in fitness effects from two or more mutations, termed epistasis, can result in compensation of deleterious mutations or negation of beneficial mutations. Recent evidence shows the importance of epistasis in individual evolutionary pathways. However, an unresolved question in molecular evolution is how often and how significantly fitness effects change in alternative genetic backgrounds.
Results: To answer this question, we quantified the effects of all single mutations and double mutations between all positions in the IgG-binding domain of protein G (GB1). By observing the first two steps of all possible evolutionary pathways using this fitness profile, we were able to characterize the extent and magnitude of pairwise epistasis throughout an entire protein molecule. Furthermore, we developed a novel approach to quantitatively determine the effects of single mutations on structural stability (Delta Delta G(U)). This enabled determination of the importance of stability effects in functional epistasis.
Conclusions: Our results illustrate common biophysical mechanisms for occurrences of positive and negative epistasis. Our results show pervasive positive epistasis within a conformationally dynamic network of residues. The stability analysis shows that significant negative epistasis, which is more common than positive epistasis, mostly occurs between combinations of destabilizing mutations. Furthermore, we show that although significant positive epistasis is rare, many deleterious mutations are beneficial in at least one alternative mutational background. The distribution of conditionally beneficial mutations throughout the domain demonstrates that the functional portion of sequence space can be significantly expanded by epistasis.
RI Wu, Nicholas/H-3822-2015
OI Wu, Nicholas/0000-0002-9078-6697
SN 0960-9822
EI 1879-0445
PD NOV 17
PY 2014
VL 24
IS 22
BP 2643
EP 2651
DI 10.1016/j.cub.2014.09.072
UT WOS:000345189700017
PM 25455030
ER

PT J
AU Kanamori, T
Fujino, Y
Ueda, T
AF Kanamori, Takashi
Fujino, Yasuhiro
Ueda, Takuya
TI PURE ribosome display and its application in antibody technology
SO BIOCHIMICA ET BIOPHYSICA ACTA-PROTEINS AND PROTEOMICS
AB Ribosome display utilizes formation of the mRNA-ribosome-polypeptide ternary complex in a cell-free protein synthesis system to link genotype (mRNA) to phenotype (polypeptide). However, the presence of intrinsic components, such as nucleases in the cell-extract-based cell-free protein synthesis system, reduces the stability of the ternary complex, which would prevent attainment of reliable results. We have developed an efficient and highly controllable ribosome display system using the PURE (Protein synthesis Using Recombinant Elements) system. The mRNA-ribosome-polypeptide ternary complex is highly stable in the PURE system, and the selected mRNA can be easily recovered because activities of nucleases and other inhibitory factors are very low in the PURE system. We have applied the PURE ribosome display to antibody engineering approaches, such as epitope mapping and affinity maturation of antibodies, and obtained results showing that the PURE ribosome display is more efficient than the conventional method. We believe that the PURE ribosome display can contribute to the development of useful antibodies. This article is part of a Special Issue entitled: Recent advances in molecular engineering of antibody.
RI Ueda, Takuya/K-5217-2014
OI Ueda, Takuya/0000-0002-7760-8271
SN 1570-9639
EI 0006-3002
PD NOV
PY 2014
VL 1844
IS 11
SI SI
BP 1925
EP 1932
DI 10.1016/j.bbapap.2014.04.007

UT WOS:000343624200005
ER

PT J
AU Boucher, JI
Cote, P
Flynn, J
Jiang, L
Laban, A
Mishra, P
Roscoe, BP
Bolon, DNA
AF Boucher, Jeffrey I.
Cote, Pamela
Flynn, Julia
Jiang, Li
Laban, Aneth
Mishra, Parul
Roscoe, Benjamin P.
Bolon, Daniel N. A.
TI Viewing Protein Fitness Landscapes Through a Next-Gen Lens
SO GENETICS
AB High-throughput sequencing has enabled many powerful approaches in biological research. Here, we review sequencing approaches to measure frequency changes within engineered mutational libraries subject to selection. These analyses can provide direct estimates of biochemical and fitness effects for all individual mutations across entire genes (and likely compact genomes in the near future) in genetically tractable systems such as microbes, viruses, and mammalian cells. The effects of mutations on experimental fitness can be assessed using sequencing to monitor time-dependent changes in mutant frequency during bulk competitions. The impact of mutations on biochemical functions can be determined using reporters or other means of separating variants based on individual activities (e.g., binding affinity for a partner molecule can be interrogated using surface display of libraries of mutant proteins and isolation of bound and unbound populations). The comprehensive investigation of mutant effects on both biochemical function and experimental fitness provide promising new avenues to investigate the connections between biochemistry, cell physiology, and evolution. We summarize recent findings from systematic mutational analyses; describe how they relate to a field rich in both theory and experimentation; and highlight how they may contribute to ongoing and future research into protein structure-function relationships, systems-level descriptions of cell physiology, and population-genetic inferences on the relative contributions of selection and drift.
OI Bolon, Daniel/0000-0001-5857-6676
SN 0016-6731
EI 1943-2631
PD OCT
PY 2014
VL 198
IS 2
BP 461
EP 471
DI 10.1534/genetics.114.168351
UT WOS:000343885300012
PM 25316787
ER

PT J
AU Bloom, JD
AF Bloom, Jesse D.
TI An Experimentally Informed Evolutionary Model Improves Phylogenetic Fit to Divergent Lactamase Homologs
SO MOLECULAR BIOLOGY AND EVOLUTION
AB Phylogenetic analyses of molecular data require a quantitative model for how sequences evolve. Traditionally, the details of the site-specific selection that governs sequence evolution are not known a priori, making it challenging to create evolutionary models that adequately capture the heterogeneity of selection at different sites. However, recent advances in high-throughput experiments have made it possible to quantify the effects of all single mutations on gene function. I have previously shown that such high-throughput experiments can be combined with knowledge of underlying mutation rates to create a parameter-free evolutionary model that describes the phylogeny of influenza nucleoprotein far better than commonly used existing models. Here, I extend this work by showing that published experimental data on TEM-1 beta-lactamase (Firnberg E, Labonte JW, Gray JJ, Ostermeier M. 2014. A comprehensive, high-resolution map of a gene's fitness landscape. Mol Biol Evol. 31:1581-1592) can be combined with a few mutation rate parameters to create an evolutionary model that describes beta-lactamase phylogenies much better than most common existing models. This experimentally informed evolutionary model is superior even for homologs that are substantially diverged (about 35% divergence at the protein level) from the TEM-1 parent that was the subject of the experimental study. These results suggest that experimental measurements can inform phylogenetic evolutionary models that are applicable to homologs that span a substantial range of sequence divergence.
RI Bloom, Jesse/C-6837-2013
OI Bloom, Jesse/0000-0003-1267-3408
SN 0737-4038

EI 1537-1719
PD OCT
PY 2014
VL 31
IS 10
BP 2753
EP 2769
DI 10.1093/molbev/msu220
UT WOS:000343402200018
PM 25063439
ER

PT J
AU Findlay, GM
Boyle, EA
Hause, RJ
Klein, JC
Shendure, J
AF Findlay, Gregory M.
Boyle, Evan A.
Hause, Ronald J.
Klein, Jason C.
Shendure, Jay
TI Saturation editing of genomic regions by multiplex homology-directed repair
SO NATURE
AB Saturation mutagenesis(1,2)-coupled to an appropriate biological assay-represents a fundamental means of achieving a high-resolution understanding of regulatory(3) and protein-coding(4) nucleic acid sequences of interest. However, mutagenized sequences introduced in trans on episomes or via random or "safe-harbour'' integration fail to capture the native context of the endogenous chromosomal locus(5). This shortcoming markedly limits the interpretability of the resulting measurements of mutational impact. Here, we couple CRISPR/Cas9 RNA-guided cleavage(6) with multiplex homology-directed repair using a complex library of donor templates to demonstrate saturation editing of genomic regions. In exon 18 of BRCA1, we replace a six-base-pair (bp) genomic region with all possible hexamers, or the full exon with all possible single nucleotide variants (SNVs), and measure strong effects on transcript abundance attributable to nonsense-mediated decay and exonic splicing elements. We similarly perform saturation genome editing of a well-conserved coding region of an essential gene, DBR1, and measure relative effects on growth that correlate with functional impact. Measurement of the functional consequences of large numbers of mutations with saturation genome editing will potentially facilitate high-resolution functional dissection of both cis-regulatory elements and trans-acting factors, as well as the interpretation of variants of uncertain significance observed in clinical sequencing.
OI Shendure, Jay/0000-0002-1516-1865; Boyle, Evan/0000-0003-4494-9771
SN 0028-0836
EI 1476-4687
PD SEP 4
PY 2014
VL 513
IS 7516
BP 120
EP +
DI 10.1038/nature13695
UT WOS:000341174800042
PM 25141179
ER

PT J
AU Fowler, DM
Stephany, JJ
Fields, S
AF Fowler, Douglas M.
Stephany, Jason J.
Fields, Stanley
TI Measuring the activity of protein variants on a large scale using deep mutational scanning
SO NATURE PROTOCOLS
AB Deep mutational scanning marries selection for protein function to high-throughput DNA sequencing in order to quantify the activity of variants of a protein on a massive scale. First, an appropriate selection system for the protein function of interest is identified and validated. Second, a library of variants is created, introduced into the selection system and subjected to selection. Third, library DNA is recovered throughout the selection and deep-sequenced. Finally, a functional score for each variant is calculated on the basis of the change in the frequency of the variant during the selection. This protocol describes the steps that must be carried out to generate a large-scale mutagenesis data set consisting of functional scores for up to hundreds of thousands of variants of a protein of interest. Establishing an assay, generating a library of variants and carrying out a selection and its accompanying sequencing takes on the order of 4-6 weeks; the initial data analysis can be completed in 1 week.

SN 1754-2189
EI 1750-2799
PD SEP
PY 2014
VL 9
IS 9
BP 2267
EP 2284
DI 10.1038/nprot.2014.153
UT WOS:000343227100021
PM 25167058
ER

PT J
AU Bloom, JD
AF Bloom, Jesse D.
TI An Experimentally Determined Evolutionary Model Dramatically Improves Phylogenetic Fit
SO MOLECULAR BIOLOGY AND EVOLUTION
AB All modern approaches to molecular phylogenetics require a quantitative model for how genes evolve. Unfortunately, existing evolutionary models do not realistically represent the site-heterogeneous selection that governs actual sequence change. Attempts to remedy this problem have involved augmenting these models with a burgeoning number of free parameters. Here, I demonstrate an alternative: Experimental determination of a parameter-free evolutionary model via mutagenesis, functional selection, and deep sequencing. Using this strategy, I create an evolutionary model for influenza nucleoprotein that describes the gene phylogeny far better than existing models with dozens or even hundreds of free parameters. Emerging high-throughput experimental strategies such as the one employed here provide fundamentally new information that has the potential to transform the sensitivity of phylogenetic and genetic analyses.
RI Bloom, Jesse/C-6837-2013
OI Bloom, Jesse/0000-0003-1267-3408
SN 0737-4038
EI 1537-1719
PD AUG
PY 2014
VL 31
IS 8
BP 1956
EP 1978
DI 10.1093/molbev/msu173
UT WOS:000339927800002
PM 24859245
ER

PT J
AU Fowler, DM
Fields, S
AF Fowler, Douglas M.
Fields, Stanley
TI Deep mutational scanning: a new style of protein science
SO NATURE METHODS
AB Mutagenesis Provides insight into proteins, but only recently have assays that couple genotype to phenotype been used to assess the activities of as many as 1 million mutant versions of a protein in a single experiment. This approach-'deep mutational scanning'-yields large-scale data sets that can reveal intrinsic protein properties, protein behavior within cells and the consequences of human genetic variation. Deep mutational scanning is transforming the study of proteins, but many challenges must be tackled for it to fulfill its promise.
SN 1548-7091
EI 1548-7105
PD AUG
PY 2014
VL 11
IS 8
BP 801
EP 807
DI 10.1038/NMETH.3027
UT WOS:000340075600018
PM 25075907
ER

PT J
AU Roscoe, BP

Bolon, DNA
AF Roscoe, Benjamin P.
Bolon, Daniel N. A.
TI Systematic Exploration of Ubiquitin Sequence, E1 Activation Efficiency, and Experimental Fitness in Yeast
SO JOURNAL OF MOLECULAR BIOLOGY
AB The complexity of biological interaction networks poses a challenge to understanding the function of individual connections in the overall network. To address this challenge, we developed a high-throughput reverse engineering strategy to analyze how thousands of specific perturbations (encompassing all point mutations in a central gene) impact both a specific edge (interaction to a directly connected node) and an overall network function. We analyzed the effects of ubiquitin mutations on activation by the E1 enzyme and compared these to effects on yeast growth rate. Using this approach, we delineated ubiquitin mutations that selectively impacted the ubiquitin-E1 edge. We find that the elasticity function relating the efficiency of ubiquitin-E1 interaction to growth rate is non-linear and that a greater than 50-fold decrease in E1 activation efficiency is required to reduce growth rate by 2-fold. Despite the robustness of fitness to decreases in E1 activation efficiency, the effects of most ubiquitin mutations on El activation paralleled the effects on growth rate. Our observations indicate that most ubiquitin mutations that disrupt E1 activation also disrupt other functions. The structurally characterized ubiquitin-E1 interlace encompasses the interlaces of ubiquitin with most other known binding partners, and we propose that this enables E1 in wild-type cells to selectively activate ubiquitin protein molecules capable of binding to other partners from the cytoplasmic pool of ubiquitin protein that will include molecules with chemical damage and/or errors from transcription and translation. (C) 2014 Elsevier Ltd. All rights reserved.
OI Bolon, Daniel/0000-0001-5857-6676
SN 0022-2836
EI 1089-8638
PD JUL 29
PY 2014
VL 426
IS 15
BP 2854
EP 2870
DI 10.1016/j.jmb.2014.05.019
UT WOS:000340302700012
PM 24862281
ER

PT J
AU Thyagarajan, B
Bloom, JD
AF Thyagarajan, Bargavi
Bloom, Jesse D.
TI The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin
SO ELIFE
AB Influenza is notable for its evolutionary capacity to escape immunity targeting the viral hemagglutinin. We used deep mutational scanning to examine the extent to which a high inherent mutational tolerance contributes to this antigenic evolvability. We created mutant viruses that incorporate most of the approximate to 10(4) amino-acid mutations to hemagglutinin from A/WSN/1933 (H1N1) influenza. After passaging these viruses in tissue culture to select for functional variants, we used deep sequencing to quantify mutation frequencies before and after selection. These data enable us to infer the preference for each amino acid at each site in hemagglutinin. These inferences are consistent with existing knowledge about the protein's structure and function, and can be used to create a model that describes hemagglutinin's evolution far better than existing phylogenetic models. We show that hemagglutinin has a high inherent tolerance for mutations at antigenic sites, suggesting that this is one factor contributing to influenza's antigenic evolution.
SN 2050-084X
PD JUL 8
PY 2014
VL 3
AR e03300
DI 10.7554/eLife.03300
UT WOS:000209690500001
PM 25006036
ER

PT J
AU Otwinowski, J
Plotkin, JB
AF Otwinowski, Jakub
Plotkin, Joshua B.
TI Inferring fitness landscapes by regression produces biased estimates of epistasis
SO PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA
AB The genotype-fitness map plays a fundamental role in shaping the dynamics of evolution. However, it is difficult to directly measure a fitness

landscape in practice, because the number of possible genotypes is astronomical. One approach is to sample as many genotypes as possible, measure their fitnesses, and fit a statistical model of the landscape that includes additive and pairwise interactive effects between loci. Here, we elucidate the pitfalls of using such regressions by studying artificial but mathematically convenient fitness landscapes. We identify two sources of bias inherent in these regression procedures, each of which tends to underestimate high fitnesses and overestimate low fitnesses. We characterize these biases for random sampling of genotypes as well as samples drawn from a population under selection in the Wright-Fisher model of evolutionary dynamics. We show that common measures of epistasis, such as the number of monotonically increasing paths between ancestral and derived genotypes, the prevalence of sign epistasis, and the number of local fitness maxima, are distorted in the inferred landscape. As a result, the inferred landscape will provide systematically biased predictions for the dynamics of adaptation. We identify the same biases in a computational RNA-folding landscape as well as regulatory sequence binding data treated with the same fitting procedure. Finally, we present a method to ameliorate these biases in some cases.
RI Otwinowski, Jakub/B-1289-2011
OI Otwinowski, Jakub/0000-0002-0341-8790
SN 0027-8424
PD JUN 3
PY 2014
VL 111
IS 22
BP E2301
EP E2309
DI 10.1073/pnas.1400849111
UT WOS:000336687900011
PM 24843135
ER

PT J
AU Shin, H
  Cho, Y
  Choe, DH
  Jeong, Y
  Cho, S
  Kim, SC
  Cho, BK
AF Shin, HyeonSeok
  Cho, Yoobok
  Choe, Dong-hui
  Jeong, Yujin
  Cho, Suhyung
  Kim, Sun Chang
  Cho, Byung-Kwan
TI Exploring the Functional Residues in a Flavin-Binding Fluorescent Protein Using Deep Mutational Scanning
SO PLOS ONE
AB Flavin mononucleotide (FMN)-based fluorescent proteins are versatile reporters that can monitor various cellular processes in both aerobic and anaerobic conditions. However, the understanding of the role of individual amino acid residues on the protein function has been limited and has restricted the development of better functional variants. Here we examine the functional amino acid residues of Escherichia coli flavin mononucleotide binding fluorescent protein (EcFbFP) using the application of high-throughput sequencing of functional variants, termed deep mutational scanning. The variants were classified into 329 function-retained (FR) and 259 function-loss (FL) mutations, and further the mutational enrichment in each amino acid residues was weighed to find the functionally important residues of EcFbFP. We show that the crucial amino acid residues of EcFbFP lie among the FMN-binding pocket, turns and loops of the protein where conformation changes occur, and spatially clustered residues near the E56-K97 salt bridges. In addition, the mutational sensitivity of the critical residues was confirmed by site-directed mutagenesis. The deep mutational scanning of EcFbFP has demonstrated important implications for constructing better functioning protein variants.
RI Cho, Byung-Kwan/C-1830-2011; Kim, Sun Chang/C-2026-2011
SN 1932-6203
PD JUN 2
PY 2014
VL 9
IS 6
AR e97817
DI 10.1371/journal.pone.0097817
UT WOS:000336956300023
PM 24887409
ER

PT J
AU Firnberg, E
  Labonte, JW
  Gray, JJ
  Ostermeier, M

AF Firnberg, Elad
Labonte, Jason W.
Gray, Jeffrey J.
Ostermeier, Marc
TI A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape
SO MOLECULAR BIOLOGY AND EVOLUTION
AB Mutations are central to evolution, providing the genetic variation upon which selection acts. A mutation's effect on the suitability of a gene to perform a particular function (gene fitness) can be positive, negative, or neutral. Knowledge of the distribution of fitness effects (DFE) of mutations is fundamental for understanding evolutionary dynamics, molecular-level genetic variation, complex genetic disease, the accumulation of deleterious mutations, and the molecular clock. We present comprehensive DFEs for point and codon mutants of the Escherichia coli TEM-1 beta-lactamase gene and missense mutations in the TEM-1 protein. These DFEs provide insight into the inherent benefits of the genetic code's architecture, support for the hypothesis that mRNA stability dictates codon usage at the beginning of genes, an extensive framework for understanding protein mutational tolerance, and evidence that mutational effects on protein thermodynamic stability shape the DFE. Contrary to prevailing expectations, we find that deleterious effects of mutation primarily arise from a decrease in specific protein activity and not cellular protein levels.
SN 0737-4038
EI 1537-1719
PD JUN
PY 2014
VL 31
IS 6
BP 1581
EP 1592
DI 10.1093/molbev/msu081
UT WOS:000337067400024
PM 24567513
ER

PT J
AU Kosuri, S
Church, GM
AF Kosuri, Sriram
Church, George M.
TI Large-scale de novo DNA synthesis: technologies and applications
SO NATURE METHODS
AB For over 60 years, the synthetic production of new DNA sequences has helped researchers understand and engineer biology. Here we summarize methods and caveats for the de novo synthesis of DNA, with particular emphasis on recent technologies that allow for large-scale and low-cost production. In addition, we discuss emerging applications enabled by large-scale de novo DNA constructs, as well as the challenges and opportunities that lie ahead.
OI Kosuri, Sriram/0000-0002-4661-0600
SN 1548-7091
EI 1548-7105
PD MAY
PY 2014
VL 11
IS 5
BP 499
EP 507
DI 10.1038/NMETH.2918
UT WOS:000335873400014
PM 24781323
ER

PT J
AU Chrystoja, CC
Diamandis, EP
AF Chrystoja, Caitlin C.
Diamandis, Eleftherios P.
TI Whole Genome Sequencing as a Diagnostic Test: Challenges and Opportunities
SO CLINICAL CHEMISTRY
AB BACKGROUND: Extraordinary technological advances and decreases in the cost of DNA sequencing have made the possibility of whole genome sequencing (WGS) as a highly accessible clinical test for numerous indications feasible. There have been many recent, successful applications of WGS in establishing the etiology of complex diseases and guiding therapeutic decision-making in neoplastic and nonneoplastic diseases and in various aspects of reproductive health. However, there are major, but not insurmountable, obstacles to the increased clinical implementation of WGS, such as hidden costs, issues surrounding sequencing and analysis, quality assurance and standardization protocols, ethical dilemmas, and difficulties with interpretation of the results.
CONTENT: The widespread use of WGS in routine clinical practice remains a distant proposition. Prospective trials will be needed to establish if, and for

whom, the benefits of WGS will outweigh the likely substantial costs associated with follow-up tests, the risks of over-diagnosis and overtreatment, and the associated emotional distress.

SUMMARY: WGS should be carefully implemented in the clinic to allow the realization of its potential to improve patient health in specific indications. To minimize harm the use of WGS for all other reasons must be carefully evaluated before clinical implementation. (C) 2013 American Association for Clinical Chemistry

OI Chrystoja, Caitlin/0000-0001-7072-3173
SN 0009-9147
EI 1530-8561
PD MAY
PY 2014
VL 60
IS 5
BP 724
EP 733
DI 10.1373/clinchem.2013.209213
UT WOS:000335147700008
PM 24227285
ER

PT J
AU Qi, HF
Olson, CA
Wu, NC
Ke, RA
Loverdo, C
Chu, V
Truong, S
Remenyi, R
Chen, ZG
Du, YS
Su, SY
Al-Mawsawi, LQ
Wu, TT
Chen, SH
Lin, CY
Zhong, WD
Lloyd-Smith, JO
Sun, R
AF Qi, Hangfei
Olson, C. Anders
Wu, Nicholas C.
Ke, Ruian
Loverdo, Claude
Chu, Virginia
Truong, Shawna
Remenyi, Roland
Chen, Zugen
Du, Yushen
Su, Sheng-Yao
Al-Mawsawi, Laith Q.
Wu, Ting-Ting
Chen, Shu-Hua
Lin, Chung-Yen
Zhong, Weidong
Lloyd-Smith, James O.
Sun, Ren
TI A Quantitative High-Resolution Genetic Profile Rapidly Identifies Sequence Determinants of Hepatitis C Viral Fitness and Drug Sensitivity
SO PLOS PATHOGENS
AB Widely used chemical genetic screens have greatly facilitated the identification of many antiviral agents. However, the regions of interaction and inhibitory mechanisms of many therapeutic candidates have yet to be elucidated. Previous chemical screens identified Daclatasvir (BMS-790052) as a potent nonstructural protein 5A (NS5A) inhibitor for Hepatitis C virus (HCV) infection with an unclear inhibitory mechanism. Here we have developed a quantitative high-resolution genetic (qHRG) approach to systematically map the drug-protein interactions between Daclatasvir and NS5A and profile genetic barriers to Daclatasvir resistance. We implemented saturation mutagenesis in combination with next-generation sequencing technology to systematically quantify the effect of every possible amino acid substitution in the drug-targeted region (domain IA of NS5A) on replication fitness and sensitivity to Daclatasvir. This enabled determination of the residues governing drug-protein interactions. The relative fitness and drug sensitivity profiles also provide a comprehensive reference of the genetic barriers for all possible single amino acid changes during viral evolution, which we utilized to

predict clinical outcomes using mathematical models. We envision that this high-resolution profiling methodology will be useful for next-generation drug development to select drugs with higher fitness costs to resistance, and also for informing the rational use of drugs based on viral variant spectra from patients.
RI Lloyd-Smith, James/K-4080-2012; Wu, Nicholas/H-3822-2015;
OI Lloyd-Smith, James/0000-0001-7941-502X; Wu,
Nicholas/0000-0002-9078-6697; Loverdo, Claude/0000-0002-0888-1717
SN 1553-7366
EI 1553-7374
PD APR
PY 2014
VL 10
IS 4
AR e1004064
DI 10.1371/journal.ppat.1004064
UT WOS:000342033600030
PM 24722365
ER

PT J
AU Jain, PC
Varadarajan, R
AF Jain, Pankaj C.
Varadarajan, Raghavan
TI A rapid, efficient, and economical inverse polymerase chain reaction-based method for generating a site saturation mutant library
SO ANALYTICAL BIOCHEMISTRY
AB With the development of deep sequencing methodologies, it has become important to construct site saturation mutant (SSM) libraries in which every nucleotide/codon in a gene is individually randomized. We describe methodologies for the rapid, efficient, and economical construction of such libraries using inverse polymerase chain reaction (PCR). We show that if the degenerate codon is in the middle of the mutagenic primer, there is an inherent PCR bias due to the thermodynamic mismatch penalty, which decreases the proportion of unique mutants. Introducing a nucleotide bias in the primer can alleviate the problem. Alternatively, if the degenerate codon is placed at the 5' end, there is no PCR bias, which results in a higher proportion of unique mutants. This also facilitates detection of deletion mutants resulting from errors during primer synthesis. This method can be used to rapidly generate SSM libraries for any gene or nucleotide sequence, which can subsequently be screened and analyzed by deep sequencing. (C) 2013 Elsevier Inc. All rights reserved.
SN 0003-2697
EI 1096-0309
PD MAR 15
PY 2014
VL 449
BP 90
EP 98
DI 10.1016/j.ab.2013.12.002
UT WOS:000332816100013
PM 24333246
ER

PT J
AU Liachko, I
Youngblood, RA
Tsui, K
Bubb, KL
Queitsch, C
Raghuraman, MK
Nislow, C
Brewer, BJ
Dunham, MJ
AF Liachko, Ivan
Youngblood, Rachel A.
Tsui, Kyle
Bubb, Kerry L.
Queitsch, Christine
Raghuraman, M. K.
Nislow, Corey
Brewer, Bonita J.
Dunham, Maitreya J.
TI GC-Rich DNA Elements Enable Replication Origin Activity in the Methylotrophic Yeast Pichia pastoris

Dunham, Maitreya J.
TI GC-Rich DNA Elements Enable Replication Origin Activity in the Methylotrophic Yeast Pichia pastoris
SO PLOS GENETICS
AB The well-studied DNA replication origins of the model budding and fission yeasts are A/T-rich elements. However, unlike their yeast counterparts, both plant and metazoan origins are G/C-rich and are associated with transcription start sites. Here we show that an industrially important methylotrophic budding yeast, Pichia pastoris, simultaneously employs at least two types of replication origins-a G/C-rich type associated with transcription start sites and an A/T-rich type more reminiscent of typical budding and fission yeast origins. We used a suite of massively parallel sequencing tools to map and dissect P. pastoris origins comprehensively, to measure their replication dynamics, and to assay the global positioning of nucleosomes across the genome. Our results suggest that some functional overlap exists between promoter sequences and G/C-rich replication origins in P. pastoris and imply an evolutionary bifurcation of the modes of replication initiation.
OI Dunham, Maitreya/0000-0001-9944-2666; Nislow, Corey/0000-0002-4016-8874
SN 1553-7390
EI 1553-7404
PD MAR
PY 2014
VL 10
IS 3
AR e1004169
DI 10.1371/journal.pgen.1004169
UT WOS:000337144700011
PM 24603708
ER

PT J
AU Bank, C
Hietpas, RT
Wong, A
Bolon, DN
Jensen, JD
AF Bank, Claudia
Hietpas, Ryan T.
Wong, Alex
Bolon, Daniel N.
Jensen, Jeffrey D.
TI A Bayesian MCMC Approach to Assess the Complete Distribution of Fitness Effects of New Mutations: Uncovering the Potential for Adaptive Walks in Challenging Environments
SO GENETICS
AB The role of adaptation in the evolutionary process has been contentious for decades. At the heart of the century-old debate between neutralists and selectionists lies the distribution of fitness effects (DFE)that is, the selective effect of all mutations. Attempts to describe the DFE have been varied, occupying theoreticians and experimentalists alike. New high-throughput techniques stand to make important contributions to empirical efforts to characterize the DFE, but the usefulness of such approaches depends on the availability of robust statistical methods for their interpretation. We here present and discuss a Bayesian MCMC approach to estimate fitness from deep sequencing data and use it to assess the DFE for the same 560 point mutations in a coding region of Hsp90 in Saccharomyces cerevisiae across six different environmental conditions. Using these estimates, we compare the differences in the DFEs resulting from mutations covering one-, two-, and three-nucleotide steps from the wild typeshowing that multiple-step mutations harbor more potential for adaptation in challenging environments, but also tend to be more deleterious in the standard environment. All observations are discussed in the light of expectations arising from Fisher's geometric model.
OI Bank, Claudia/0000-0003-4730-758X; Bolon, Daniel/0000-0001-5857-6676
SN 1943-2631
PD MAR
PY 2014
VL 196
IS 3
BP 841
EP +
DI 10.1534/genetics.113.156190
UT WOS:000333905500020
PM 24398421
ER

PT J
AU Tripathi, A
Varadarajan, R
AF Tripathi, Arti
Varadarajan, Raghavan
TI Residue specific contributions to stability and activity inferred from

Varadarajan, R
AF Tripathi, Arti
Varadarajan, Raghavan
TI Residue specific contributions to stability and activity inferred from saturation mutagenesis and deep sequencing
SO CURRENT OPINION IN STRUCTURAL BIOLOGY
AB Multiple methods currently exist for rapid construction and screening of single-site saturation mutagenesis (SSM) libraries in which every codon or nucleotide in a DNA fragment is individually randomized. Nucleotide sequences of each library member before and after screening or selection can be obtained through deep sequencing. The relative enrichment of each mutant at each position provides information on its contribution to protein activity or ligand-binding under the conditions of the screen. Such saturation scans have been applied to diverse proteins to delineate hot-spot residues, stability determinants, and for comprehensive fitness estimates. The data have been used to design proteins with enhanced stability, activity and altered specificity relative to wild-type, to test computational predictions of binding affinity, and for protein model discrimination. Future improvements in deep sequencing read lengths and accuracy should allow comprehensive studies of epistatic effects, of combinational variation at multiple sites, and identification of spatially proximate residues.
SN 0959-440X
EI 1879-033X
PD FEB
PY 2014
VL 24
BP 63
EP 71
DI 10.1016/j.sbi.2013.12.001
UT WOS:000335100300009
PM 24721454
ER

PT J
AU Strauch, EM
Fleishman, SJ
Baker, D
AF Strauch, Eva-Maria
Fleishman, Sarel J.
Baker, David
TI Computational design of a pH-sensitive IgG binding protein
SO PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA
AB Computational design provides the opportunity to program protein-protein interactions for desired applications. We used de novo protein interface design to generate a pH-dependent Fc domain binding protein that buries immunoglobulin G (IgG) His-433. Using next-generation sequencing of naive and selected pools of a library of design variants, we generated a molecular footprint of the designed binding surface, confirming the binding mode and guiding further optimization of the balance between affinity and pH sensitivity. In biolayer interferometry experiments, the optimized design binds IgG with a K-d of similar to 4 nM at pH 8.2, and approximately 500-fold more weakly at pH 5.5. The protein is extremely stable, heat-resistant and highly expressed in bacteria, and allows pH-based control of binding for IgG affinity purification and diagnostic devices.
RI Baker, David/K-8941-2012;
OI Baker, David/0000-0001-7896-6217; Fleishman, Sarel/0000-0003-3177-7560
SN 0027-8424
PD JAN 14
PY 2014
VL 111
IS 2
BP 675
EP 680
DI 10.1073/pnas.1313605111
UT WOS:000329614500035
PM 24381156
ER

PT J
AU Chang, HJ
Jian, JW
Hsu, HJ
Lee, YC
Chen, HS
You, JJ
Hou, SC
Shao, CY
Chen, YJ
Chiu, KP

Peng, HP
Lee, KH
Yang, AS
AF Chang, Hung-Ju
Jian, Jhih-Wei
Hsu, Hung-Ju
Lee, Yu-Ching
Chen, Hong-Sen
You, Jhong-Jhe
Hou, Shin-Chen
Shao, Chih-Yun
Chen, Yen-Ju
Chiu, Kuo-Ping
Peng, Hung-Pin
Lee, Kuo Hao
Yang, An-Suei
TI Loop-Sequence Features and Stability Determinants in Antibody Variable Domains by High-Throughput Experiments
SO STRUCTURE
AB Protein loops are frequently considered as critical determinants in protein structure and function. Recent advances in high-throughput methods for DNA sequencing and thermal stability measurement have enabled effective exploration of sequence-structure-function relationships in local protein regions. Using these data-intensive technologies, we investigated the sequence-structure-function relationships of six complementarity-determining regions (CDRs) and ten non-CDR loops in the variable domains of a model vascular endothelial growth factor (VEGF)-binding single-chain antibody variable fragment (scFv) whose sequence had been optimized via a consensus-sequence approach. The results show that only a handful of residues involving long-range tertiary interactions distant from the antigen-binding site are strongly coupled with antigen binding. This implies that the loops are passive regions in protein folding; the essential sequences of these regions are dictated by conserved tertiary interactions and the consensus local loop-sequence features contribute little to protein stability and function.
SN 0969-2126
EI 1878-4186
PD JAN 7
PY 2014
VL 22
IS 1
BP 9
EP 21
DI 10.1016/j.str.2013.10.005
UT WOS:000329593000004
PM 24268648
ER

PT J
AU Hsu, HJ
Lee, KH
Jian, JW
Chang, HJ
Yu, CM
Lee, YC
Chen, IC
Peng, HP
Wu, CY
Huang, YF
Shao, CY
Chiu, KP
Yang, AS
AF Hsu, Hung-Ju
Lee, Kuo Hao
Jian, Jhih-Wei
Chang, Hung-Ju
Yu, Chung-Ming
Lee, Yu-Ching
Chen, Ing-Chien
Peng, Hung-Pin
Wu, Chih Yuan
Huang, Yu-Feng
Shao, Chih-Yun
Chiu, Kuo Ping
Yang, An-Suei

Shao, Chih-Yun
Chiu, Kuo Ping
Yang, An-Suei
TI Antibody Variable Domain Interface and Framework Sequence Requirements for Stability and Function by High-Throughput Experiments
SO STRUCTURE
AB Protein structural stability and biological functionality are dictated by the formation of intradomain cores and interdomain interfaces, but the intricate sequence-structure-function interrelationships in the packing of protein cores and interfaces remain difficult to elucidate due to the intractability of enumerating all packing possibilities and assessing the consequences of all the variations. In this work, groups of beta strand residues of model antibody variable domains were randomized with saturated mutagenesis and the functional variants were selected for high-throughput sequencing and high-throughput thermal stability measurements. The results show that the sequence preferences of the intradomain hydrophobic core residues are strikingly flexible among hydrophobic residues, implying that these residues are coupled indirectly with antigen binding through energetic stabilization of the protein structures. By contrast, the interdomain interface residues are directly coupled with antigen binding. The interdomain interlace should be treated as an integral part of the antigen-binding site.
SN 0969-2126
EI 1878-4186
PD JAN 7
PY 2014
VL 22
IS 1
BP 22
EP 34
DI 10.1016/j.str.2013.10.006
UT WOS:000329593000005
PM 24268647
ER

PT J
AU Gajula, KS
Huwe, PJ
Mo, CY
Crawford, DJ
Stivers, JT
Radhakrishnan, R
Kohli, RM
AF Gajula, Kiran S.
Huwe, Peter J.
Mo, Charlie Y.
Crawford, Daniel J.
Stivers, James T.
Radhakrishnan, Ravi
Kohli, Rahul M.
TI High-throughput mutagenesis reveals functional determinants for DNA targeting by activation-induced deaminase
SO NUCLEIC ACIDS RESEARCH
AB Antibody maturation is a critical immune process governed by the enzyme activation-induced deaminase (AID), a member of the AID/APOBEC DNA deaminase family. AID/APOBEC deaminases preferentially target cytosine within distinct preferred sequence motifs in DNA, with specificity largely conferred by a small 9-11 residue protein loop that differs among family members. Here, we aimed to determine the key functional characteristics of this protein loop in AID and to thereby inform our understanding of the mode of DNA engagement. To this end, we developed a methodology (Sat-Sel-Seq) that couples saturation mutagenesis at each position across the targeting loop, with iterative functional selection and next-generation sequencing. This high-throughput mutational analysis revealed dominant characteristics for residues within the loop and additionally yielded enzymatic variants that enhance deaminase activity. To rationalize these functional requirements, we performed molecular dynamics simulations that suggest that AID and its hyperactive variants can engage DNA in multiple specific modes. These findings align with AID's competing requirements for specificity and flexibility to efficiently drive antibody maturation. Beyond insights into the AID-DNA interface, our Sat-Sel-Seq approach also serves to further expand the repertoire of techniques for deep positional scanning and may find general utility for high-throughput analysis of protein function.
OI Kohli, Rahul/0000-0002-7689-5678
SN 0305-1048
EI 1362-4962
PY 2014
VL 42
IS 15
BP 9964
EP 9975
DI 10.1093/nar/gku689
UT WOS:000343220300041
PM 25064858

ER

PT J
AU Melnikov, A
Rogov, P
Wang, L
Gnirke, A
Mikkelsen, TS
AF Melnikov, Alexandre
Rogov, Peter
Wang, Li
Gnirke, Andreas
Mikkelsen, Tarjei S.
TI Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes
SO NUCLEIC ACIDS RESEARCH
AB Deep mutational scanning has emerged as a promising tool for mapping sequence-activity relationships in proteins, ribonucleic acid and deoxyribonucleic acid. In this approach, diverse variants of a sequence of interest are first ranked according to their activities in a relevant assay, and this ranking is then used to infer the shape of the fitness landscape around the wild-type sequence. Little is currently known, however, about the degree to which such fitness landscapes are dependent on the specific assay conditions from which they are inferred. To explore this issue, we performed comprehensive single-substitution mutational scanning of APH(3')II, a Tn5 transposon-derived kinase that confers resistance to aminoglycoside antibiotics, in Escherichia coli under selection with each of six structurally diverse antibiotics at a range of inhibitory concentrations. We found that the resulting local fitness landscapes showed significant dependence on both antibiotic structure and concentration, and that this dependence can be exploited to guide protein engineering. Specifically, we found that differential analysis of fitness landscapes allowed us to generate synthetic APH(3')II variants with orthogonal substrate specificities.
SN 0305-1048
EI 1362-4962
PY 2014
VL 42
IS 14
AR e112
DI 10.1093/nar/gku511
UT WOS:000343219200003
PM 24914046
ER

PT J
AU Meinhardt, S
Manley, MW
Parente, DJ
Swint-Kruse, L
AF Meinhardt, Sarah
Manley, Michael W., Jr.
Parente, Daniel J.
Swint-Kruse, Liskin
TI Rheostats and Toggle Switches for Modulating Protein Function
SO PLOS ONE
AB The millions of protein sequences generated by genomics are expected to transform protein engineering and personalized medicine. To achieve these goals, tools for predicting outcomes of amino acid changes must be improved. Currently, advances are hampered by insufficient experimental data about nonconserved amino acid positions. Since the property "nonconserved" is identified using a sequence alignment, we designed experiments to recapitulate that context: Mutagenesis and functional characterization was carried out in 15 LacI/GalR homologs (rows) at 12 nonconserved positions (columns). Multiple substitutions were made at each position, to reveal how various amino acids of a nonconserved column were tolerated in each protein row. Results showed that amino acid preferences of nonconserved positions were highly context-dependent, had few correlations with physico-chemical similarities, and were not predictable from their occurrence in natural LacI/GalR sequences. Further, unlike the "toggle switch" behaviors of conserved positions, substitutions at nonconserved positions could be rank-ordered to show a "rheostatic", progressive effect on function that spanned several orders of magnitude. Comparisons to various sequence analyses suggested that conserved and strongly co-evolving positions act as functional toggles, whereas other important, nonconserved positions serve as rheostats for modifying protein function. Both the presence of rheostat positions and the sequence analysis strategy appear to be generalizable to other protein families and should be considered when engineering protein modifications or predicting the impact of protein polymorphisms.
SN 1932-6203
PD DEC 30
PY 2013
VL 8
IS 12
AR e83502
DI 10.1371/journal.pone.0083502
UT WOS:000329194700047

PM 24386217
ER

PT J
AU Moal, IH
  Moretti, R
  Baker, D
  Fernandez-Recio, J
AF Moal, Iain H.
  Moretti, Rocco
  Baker, David
  Fernandez-Recio, Juan
TI Scoring functions for protein-protein interactions
SO CURRENT OPINION IN STRUCTURAL BIOLOGY
AB The computational evaluation of protein-protein interactions will play an important role in organising the wealth of data being generated by high-throughput initiatives. Here we discuss future applications, report recent developments and identify areas requiring further investigation. Many functions have been developed to quantify the structural and energetic properties of interacting proteins, finding use in interrelated challenges revolving around the relationship between sequence, structure and binding free energy. These include loop modelling, side-chain refinement, docking, multimer assembly, affinity prediction, affinity change upon mutation, hotspots location and interface design. Information derived from models optimised for one of these challenges can be used to benefit the others, and can be unified within the theoretical frameworks of multi-task learning and Pareto-optimal multi-objective learning.
OI Fernandez-Recio, Juan/0000-0002-3986-7686; Moal,
  Iain/0000-0002-4960-5487
SN 0959-440X
EI 1879-033X
PD DEC
PY 2013
VL 23
IS 6
BP 862
EP 867
DI 10.1016/j.sbi.2013.06.017
UT WOS:000329148700011
PM 23871100
ER

PT J
AU Lensink, MF
  Wodak, SJ
AF Lensink, Marc F.
  Wodak, Shoshana J.
TI Docking, scoring, and affinity prediction in CAPRI
SO PROTEINS-STRUCTURE FUNCTION AND BIOINFORMATICS
AB We present the fifth evaluation of docking and related scoring methods used in the community-wide experiment on the Critical Assessment of Predicted Interactions (CAPRI). The evaluation examined predictions submitted for a total of 15 targets in eight CAPRI rounds held during the years 2010-2012. The targets represented one the most diverse set tackled by the CAPRI community so far. They included only 10 classical docking and scoring problems. In one of the classical targets, the new challenge was to predict the position of water molecules in the protein-protein interface. The remaining five targets represented other new challenges that involved estimating the relative binding affinity and the effect of point mutations on the stability of designed and natural protein-protein complexes. Although the 10 classical CAPRI targets included two difficult multicomponent systems, and a protein-oligosaccharide complex with which CAPRI participants had little experience, this evaluation indicates that the performance of docking and scoring methods has remained quite robust. More remarkably, we find that automatic docking servers exhibit a significantly improved performance, with some servers now performing on par with predictions done by humans. The performance of CAPRI participants in the new challenges, briefly reviewed here, was mediocre overall, but some groups did relatively well and their approaches suggested ways of improving methods for designing binders and for estimating the free energies of protein assemblies, which should impact the field of protein modeling and design as a whole. Proteins 2013; 81:2082-2095. (c) 2013 Wiley Periodicals, Inc.
RI Lensink, Marc/A-1678-2008
OI Lensink, Marc/0000-0003-3957-9470
SN 0887-3585
EI 1097-0134
PD DEC
PY 2013
VL 81
IS 12
SI SI
BP 2082
EP 2095
DI 10.1002/prot.24428

UT WOS:000327344300003
PM 24115211
ER

PT J
AU Johnsen, JM
Nickerson, DA
Reiner, AP
AF Johnsen, Jill M.
Nickerson, Deborah A.
Reiner, Alex P.
TI Massively parallel sequencing: the new frontier of hematologic genomics
SO BLOOD
AB Genomic technologies are becoming a routine part of human genetic analysis. The exponential growth in DNA sequencing capability has brought an unprecedented understanding of human genetic variation and the identification of thousands of variants that impact human health. In this review, we describe the different types of DNA variation and provide an overview of existing DNA sequencing technologies and their applications. As genomic technologies and knowledge continue to advance, they will become integral in clinical practice. To accomplish the goal of personalized genomic medicine for patients, close collaborations between researchers and clinicians will be essential to develop and curate deep databases of genetic variation and their associated phenotypes.
SN 0006-4971
EI 1528-0020
PD NOV 7
PY 2013
VL 122
IS 19
BP 3268
EP 3275
DI 10.1182/blood-2013-07-460287
UT WOS:000327466100011
PM 24021669
ER

PT J
AU Kim, I
Miller, CR
Young, DL
Fields, S
AF Kim, Ikjin
Miller, Christina R.
Young, David L.
Fields, Stanley
TI High-throughput Analysis of in vivo Protein Stability
SO MOLECULAR & CELLULAR PROTEOMICS
AB Determining the half-life of proteins is critical for an understanding of virtually all cellular processes. Current methods for measuring in vivo protein stability, including large-scale approaches, are limited in their throughput or in their ability to discriminate among small differences in stability. We developed a new method, Stable-seq, which uses a simple genetic selection combined with high-throughput DNA sequencing to assess the in vivo stability of a large number of variants of a protein. The variants are fused to a metabolic enzyme, which here is the yeast Leu2 protein. Plasmids encoding these Leu2 fusion proteins are transformed into yeast, with the resultant fusion proteins accumulating to different levels based on their stability and leading to different doubling times when the yeast are grown in the absence of leucine. Sequencing of an input population of variants of a protein and the population of variants after leucine selection allows the stability of tens of thousands of variants to be scored in parallel. By applying the Stable-seq method to variants of the protein degradation signal Deg1 from the yeast Mat2 protein, we generated a high-resolution map that reveals the effect of approximate to 30,000 mutations on protein stability. We identified mutations that likely affect stability by changing the activity of the degron, by leading to translation from new start codons, or by affecting N-terminal processing. Stable-seq should be applicable to other organisms via the use of suitable reporter proteins, as well as to the analysis of complex mixtures of fusion proteins.
SN 1535-9476
EI 1535-9484
PD NOV
PY 2013
VL 12
IS 11
BP 3370
EP 3378
DI 10.1074/mcp.O113.031708
UT WOS:000328816000028
PM 23897579
ER

PT J
AU Smith, JD
McManus, KF
Fraser, HB
AF Smith, Justin D.
McManus, Kimberly F.
Fraser, Hunter B.
TI A Novel Test for Selection on cis-Regulatory Elements Reveals Positive and Negative Selection Acting on Mammalian Transcriptional Enhancers
SO MOLECULAR BIOLOGY AND EVOLUTION
AB Measuring natural selection on genomic elements involved in the cis-regulation of gene expression-such as transcriptional enhancers and promoters-is critical for understanding the evolution of genomes, yet it remains a major challenge. Many studies have attempted to detect positive or negative selection in these noncoding elements by searching for those with the fastest or slowest rates of evolution, but this can be problematic. Here, we introduce a new approach to this issue, and demonstrate its utility on three mammalian transcriptional enhancers. Using results from saturation mutagenesis studies of these enhancers, we classified all possible point mutations as upregulating, downregulating, or silent, and determined which of these mutations have occurred on each branch of a phylogeny. Applying a framework analogous to K-a/K-s in protein-coding genes, we measured the strength of selection on upregulating and downregulating mutations, in specific branches as well as entire phylogenies. We discovered distinct modes of selection acting on different enhancers: although all three have experienced negative selection against downregulating mutations, the selection pressures on upregulating mutations vary. In one case, we detected positive selection for upregulation, whereas the other two had no detectable selection on upregulating mutations. Our methodology is applicable to the growing number of saturation mutagenesis data sets, and provides a detailed picture of the mode and strength of natural selection acting on cis-regulatory elements.
OI Smith, Justin/0000-0003-3079-3534
SN 0737-4038
EI 1537-1719
PD NOV
PY 2013
VL 30
IS 11
BP 2509
EP 2518
DI 10.1093/molbev/mst134
UT WOS:000326745300011
PM 23904330
ER

PT J
AU Moretti, R
Fleishman, SJ
Agius, R
Torchala, M
Bates, PA
Kastritis, PL
Rodrigues, JPGLM
Trellet, M
Bonvin, AMJJ
Cui, M
Rooman, M
Gillis, D
Dehouck, Y
Moal, I
Romero-Durana, M
Perez-Cano, L
Pallara, C
Jimenez, B
Fernandez-Recio, J
Flores, S
Pacella, M
Kilambi, KP
Gray, JJ
Popov, P
Grudinin, S
Esquivel-Rodriguez, J
Kihara, D
Zhao, N
Korkin, D
Zhu, XL
Demerdash, ONA

Mitchell, JC
Kanamori, E
Tsuchiya, Y
Nakamura, H
Lee, H
Park, H
Seok, C
Sarmiento, J
Liang, SD
Teraguchi, S
Standley, DM
Shimoyama, H
Terashi, G
Takeda-Shitaka, M
Iwadate, M
Umeyama, H
Beglov, D
Hall, DR
Kozakov, D
Vajda, S
Pierce, BG
Hwang, H
Vreven, T
Weng, ZP
Huang, YY
Li, HT
Yang, XF
Ji, XF
Liu, SY
Xiao, Y
Zacharias, M
Qin, SB
Zhou, HX
Huang, SY
Zou, XQ
Velankar, S
Janin, J
Wodak, SJ
Baker, D
AF Moretti, Rocco
Fleishman, Sarel J.
Agius, Rudi
Torchala, Mieczyslaw
Bates, Paul A.
Kastritis, Panagiotis L.
Rodrigues, Joao P. G. L. M.
Trellet, Mikael
Bonvin, Alexandre M. J. J.
Cui, Meng
Rooman, Marianne
Gillis, Dimitri
Dehouck, Yves
Moal, Iain
Romero-Durana, Miguel
Perez-Cano, Laura
Pallara, Chiara
Jimenez, Brian
Fernandez-Recio, Juan
Flores, Samuel
Pacella, Michael
Kilambi, Krishna Praneeth
Gray, Jeffrey J.
Popov, Petr
Grudinin, Sergei
Esquivel-Rodriguez, Juan
Kihara, Daisuke
Zhao, Nan

Korkin, Dmitry
Zhu, Xiaolei
Demerdash, Omar N. A.
Mitchell, Julie C.
Kanamori, Eiji
Tsuchiya, Yuko
Nakamura, Haruki
Lee, Hasup
Park, Hahnbeom
Seok, Chaok
Sarmiento, Jamica
Liang, Shide
Teraguchi, Shusuke
Standley, Daron M.
Shimoyama, Hiromitsu
Terashi, Genki
Takeda-Shitaka, Mayuko
Iwadate, Mitsuo
Umeyama, Hideaki
Beglov, Dmitri
Hall, David R.
Kozakov, Dima
Vajda, Sandor
Pierce, Brian G.
Hwang, Howook
Vreven, Thom
Weng, Zhiping
Huang, Yangyu
Li, Haotian
Yang, Xiufeng
Ji, Xiaofeng
Liu, Shiyong
Xiao, Yi
Zacharias, Martin
Qin, Sanbo
Zhou, Huan-Xiang
Huang, Sheng-You
Zou, Xiaoqin
Velankar, Sameer
Janin, Joel
Wodak, Shoshana J.
Baker, David
TI Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions
SO PROTEINS-STRUCTURE FUNCTION AND BIOINFORMATICS
AB Community-wide blind prediction experiments such as CAPRI and CASP provide an objective measure of the current state of predictive methodology. Here we describe a community-wide assessment of methods to predict the effects of mutations on protein-protein interactions. Twenty-two groups predicted the effects of comprehensive saturation mutagenesis for two designed influenza hemagglutinin binders and the results were compared with experimental yeast display enrichment data obtained using deep sequencing. The most successful methods explicitly considered the effects of mutation on monomer stability in addition to binding affinity, carried out explicit side-chain sampling and backbone relaxation, evaluated packing, electrostatic, and solvation effects, and correctly identified around a third of the beneficial mutations. Much room for improvement remains for even the best techniques, and large-scale fitness landscapes should continue to provide an excellent test bed for continued evaluation of both existing and new prediction methodologies. Proteins 2013; 81:1980-1987. (c) 2013 Wiley Periodicals, Inc.
RI Kastritis, Panagiotis/F-2498-2010; Zhou, Huan-Xiang/M-5170-2016; Gray, Jeffrey/B-5682-2008; Standley, Daron/D-2343-2009; liu, shiyong/A-9370-2011; Bonvin, Alexandre/A-5420-2009; Rodrigues, Joao/J-6579-2013; Zhao, Nan/K-4015-2015; Pallara, Chiara/F-9441-2016; Popov, Petr/G-9638-2016; Baker, David/K-8941-2012
OI Torchala, Mieczyslaw/0000-0002-4542-9156; Velankar, Sameer/0000-0002-8439-5964; Jimenez-Garcia, Brian/0000-0001-7786-2109; Dehouck, Yves/0000-0002-7401-104X; Fernandez-Recio, Juan/0000-0002-3986-7686; Moal, Iain/0000-0002-4960-5487; Fleishman, Sarel/0000-0003-3177-7560; Zhou, Huan-Xiang/0000-0001-9020-0302; Gilis, Dimitri/0000-0001-9009-9996; Kastritis, Panagiotis/0000-0002-1463-8422; Gray, Jeffrey/0000-0001-6380-2324; Bonvin, Alexandre/0000-0001-7369-1322; Rodrigues, Joao/0000-0001-9796-3193; Zhao, Nan/0000-0002-7897-2374; Pallara, Chiara/0000-0003-3005-343X;

Baker, David/0000-0001-7896-6217
SN 0887-3585
EI 1097-0134
PD NOV
PY 2013
VL 81
IS 11
BP 1980
EP 1987
DI 10.1002/prot.24356
UT WOS:000325980300011
PM 23843247
ER

PT J
AU Melamed, D
Young, DL
Gamble, CE
Miller, CR
Fields, S
AF Melamed, Daniel
Young, David L.
Gamble, Caitlin E.
Miller, Christina R.
Fields, Stanley
TI Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein
SO RNA-A PUBLICATION OF THE RNA SOCIETY
AB The RNA recognition motif (RRM) is the most common RNA-binding domain in eukaryotes. Differences in RRM sequences dictate, in part, both RNA and protein-binding specificities and affinities. We used a deep mutational scanning approach to study the sequence-function relationship of the RRM2 domain of the Saccharomyces cerevisiae poly(A)-binding protein (Pab1). By scoring the activity of more than 100,000 unique Pab1 variants, including 1246 with single amino acid substitutions, we delineated the mutational constraints on each residue. Clustering of residues with similar mutational patterns reveals three major classes, composed principally of RNA-binding residues, of hydrophobic core residues, and of the remaining residues. The first class also includes a highly conserved residue not involved in RNA binding, G150, which can be mutated to destabilize Pab1. A comparison of the mutational sensitivity of yeast Pab1 residues to their evolutionary conservation reveals that most residues tolerate more substitutions than are present in the natural sequences, although other residues that tolerate fewer substitutions may point to specialized functions in yeast. An analysis of similar to 40,000 double mutants indicates a preference for a short distance between two mutations that display an epistatic interaction. As examples of interactions, the mutations N139T, N139S, and I157L suppress other mutations that interfere with RNA binding and protein stability. Overall, this study demonstrates that living cells can be subjected to a single assay to analyze hundreds of thousands of protein variants in parallel.
SN 1355-8382
EI 1469-9001
PD NOV
PY 2013
VL 19
IS 11
BP 1537
EP 1551
DI 10.1261/rna.040709.113
UT WOS:000325813900009
PM 24064791
ER

PT J
AU Gold, MG
Gonen, T
Scott, JD
AF Gold, Matthew G.
Gonen, Tamir
Scott, John D.
TI Local cAMP signaling in disease at a glance
SO JOURNAL OF CELL SCIENCE
SN 0021-9533
EI 1477-9137
PD OCT 15
PY 2013
VL 126
IS 20

BP 4537
EP 4543
DI 10.1242/jcs.133751
UT WOS:000325803200001
PM 24124191
ER

PT J
AU Procko, E
Hedman, R
Hamilton, K
Seetharaman, J
Fleishman, SJ
Su, M
Aramini, J
Kornhaber, G
Hunt, JF
Tong, L
Montelione, GT
Baker, D
AF Procko, Erik
Hedman, Rickard
Hamilton, Keith
Seetharaman, Jayaraman
Fleishman, Sarel J.
Su, Min
Aramini, James
Kornhaber, Gregory
Hunt, John F.
Tong, Liang
Montelione, Gaetano T.
Baker, David
TI Computational Design of a Protein-Based Enzyme Inhibitor
SO JOURNAL OF MOLECULAR BIOLOGY
AB While there has been considerable progress in designing protein-protein interactions, the design of proteins that bind polar surfaces is an unmet challenge. We describe the computational design of a protein that binds the acidic active site of hen egg lysozyme and inhibits the enzyme. The design process starts with two polar amino acids that fit deep into the enzyme active site, identifies a protein scaffold that supports these residues and is complementary in shape to the lysozyme active-site region, and finally optimizes the surrounding contact surface for high-affinity binding. Following affinity maturation, a protein designed using this method bound lysozyme with low nanomolar affinity, and a combination of NMR studies, crystallography, and knockout mutagenesis confirmed the designed binding surface and orientation. Saturation mutagenesis with selection and deep sequencing demonstrated that specific designed interactions extending well beyond the centrally grafted polar residues are critical for high-affinity binding. Published by Elsevier Ltd.
RI Baker, David/K-8941-2012;
OI Baker, David/0000-0001-7896-6217; Fleishman, Sarel/0000-0003-3177-7560
SN 0022-2836
PD SEP 23
PY 2013
VL 425
IS 18
BP 3563
EP 3575
DI 10.1016/j.jmb.2013.06.035
UT WOS:000324358800024
PM 23827138
ER

PT J
AU Ghirlanda, G
AF Ghirlanda, Giovanna
TI COMPUTATIONAL BIOLOGY A recipe for ligand-binding proteins
SO NATURE
SN 0028-0836
PD SEP 12
PY 2013
VL 501
IS 7466
BP 177

EP 178
UT WOS:000324244900031
PM 24005323
ER

PT J
AU Tinberg, CE
Khare, SD
Dou, JY
Doyle, L
Nelson, JW
Schena, A
Jankowski, W
Kalodimos, CG
Johnsson, K
Stoddard, BL
Baker, D
AF Tinberg, Christine E.
Khare, Sagar D.
Dou, Jiayi
Doyle, Lindsey
Nelson, Jorgen W.
Schena, Alberto
Jankowski, Wojciech
Kalodimos, Charalampos G.
Johnsson, Kai
Stoddard, Barry L.
Baker, David
TI Computational design of ligand-binding proteins with high affinity and selectivity
SO NATURE
AB The ability to design proteins with high affinity and selectivity for any given small molecule is a rigorous test of our understanding of the physiochemical principles that govern molecular recognition. Attempts to rationally design ligand-binding proteins have met with little success, however, and the computational design of protein-small-molecule interfaces remains an unsolved problem(1). Current approaches for designing ligand-binding proteins for medical(2) and biotechnological uses rely on raising antibodies against a target antigen in immunized animals(3,4) and/or performing laboratory-directed evolution of proteins with an existing low affinity for the desired ligand(5-7), neither of which allows complete control over the interactions involved in binding. Here we describe a general computational method for designing pre-organized and shape complementary small-molecule-binding sites, and use it to generate protein binders to the steroid digoxigenin (DIG). Of seventeen experimentally characterized designs, two bind DIG; the model of the higher affinity binder has the most energetically favourable and pre-organized interface in the design set. A comprehensive binding-fitness landscape of this design, generated by library selections and deep sequencing, was used to optimize its binding affinity to a picomolar level, and X-ray co-crystal structures of two variants show atomic-level agreement with the corresponding computational models. The optimized binder is selective for DIG over the related steroids digitoxigenin, progesterone and beta-oestradiol, and this steroid binding preference can be reprogrammed by manipulation of explicitly designed hydrogen-bonding interactions. The computational design method presented here should enable the development of a new generation of biosensors, therapeutics and diagnostics.
RI johnsson, kai/P-4222-2014; Baker, David/K-8941-2012
OI johnsson, kai/0000-0002-8002-1981; Baker, David/0000-0001-7896-6217
SN 0028-0836
PD SEP 12
PY 2013
VL 501
IS 7466
BP 212
EP +
DI 10.1038/nature12443
UT WOS:000324244900039
PM 24005320
ER

PT J
AU Zayner, JP
Antoniou, C
French, AR
Hause, RJ
Sosnick, TR
AF Zayner, Josiah P.
Antoniou, Chloe
French, Alexander R.

Hause, Ronald J., Jr.
Sosnick, Tobin R.
TI Investigating Models of Protein Function and Allostery With a Widespread Mutational Analysis of a Light-Activated Protein
SO BIOPHYSICAL JOURNAL
AB To investigate the relationship between a protein's sequence and its biophysical properties, we studied the effects of more than 100 mutations in Avena sativa light-oxygen-voltage domain 2, a model protein of the Per-Arnt-Sim family. The A. sativa light oxygen voltage domain 2 undergoes a photocycle with a conformational change involving the unfolding of the terminal helices. Whereas selection studies typically search for winners in a large population and fail to characterize many sites, we characterized the biophysical consequences of mutations throughout the protein using NMR, circular dichroism, and ultraviolet/visible spectroscopy. Despite our intention to introduce highly disruptive substitutions, most had modest or no effect on function, and many could even be considered to be more photoactive. Substitutions at evolutionarily conserved sites can have minimal effect, whereas those at nonconserved positions can have large effects, contrary to the view that the effects of mutations, especially at conserved positions, are predictable. Using predictive models, we found that the effects of mutations on biophysical function and allostery reflect a complex mixture of multiple characteristics including location, character, electrostatics, and chemistry.
OI Zayner, Josiah/0000-0003-3590-3341
SN 0006-3495
PD AUG 20
PY 2013
VL 105
IS 4
BP 1027
EP 1036
DI 10.1016/j.bpj.2013.07.010
UT WOS:000323465200026
PM 23972854
ER

PT J
AU Harms, MJ
Thornton, JW
AF Harms, Michael J.
Thornton, Joseph W.
TI Evolutionary biochemistry: revealing the historical and physical causes of protein properties
SO NATURE REVIEWS GENETICS
AB The repertoire of proteins and nucleic acids in the living world is determined by evolution; their properties are determined by the laws of physics and chemistry. Explanations of these two kinds of causality - the purviews of evolutionary biology and biochemistry, respectively - are typically pursued in isolation, but many fundamental questions fall squarely at the interface of fields. Here we articulate the paradigm of evolutionary biochemistry, which aims to dissect the physical mechanisms and evolutionary processes by which biological molecules diversified and to reveal how their physical architecture facilitates and constrains their evolution. We show how an integration of evolution with biochemistry moves us towards a more complete understanding of why biological molecules have the properties that they do.
SN 1471-0056
PD AUG
PY 2013
VL 14
IS 8
BP 559
EP 571
DI 10.1038/nrg3540
UT WOS:000321956900011
PM 23864121
ER

PT J
AU Forsyth, CM
Juan, V
Akamatsu, Y
DuBridge, RB
Doan, M
Ivanov, AV
Ma, ZY
Polakoff, D
Razo, J
Wilson, K
Powers, DB
AF Forsyth, Charles M.
Juan, Veronica

Akamatsu, Yoshiko
DuBridge, Robert B.
Doan, Minhtam
Ivanov, Alexander V.
Ma, Zhiyuan
Polakoff, Dixie
Razo, Jennifer
Wilson, Keith
Powers, David B.
TI Deep mutational scanning of an antibody against epidermal growth factor receptor using mammalian cell display and massively parallel pyrosequencing
SO MABS
AB We developed a method for deep mutational scanning of antibody complementarity-determining regions (CDRs) that can determine in parallel the effect of every possible single amino acid CDR substitution on antigen binding. The method uses libraries of full length IgGs containing more than 1000 CDR point mutations displayed on mammalian cells, sorted by flow cytometry into subpopulations based on antigen affinity and analyzed by massively parallel pyrosequencing. Higher, lower and neutral affinity mutations are identified by their enrichment or depletion in the FACS subpopulations. We applied this method to a humanized version of the anti-epidermal growth factor receptor antibody cetuximab, generated a near comprehensive data set for 1060 point mutations that recapitulates previously determined structural and mutational data for these CDRs and identified 67 point mutations that increase affinity. The large-scale, comprehensive sequence-function data sets generated by this method should have broad utility for engineering properties such as antibody affinity and specificity and may advance theoretical understanding of antibody-antigen recognition.
SN 1942-0862
EI 1942-0870
PD JUL 1
PY 2013
VL 5
IS 4
BP 523
EP 532
DI 10.4161/mabs.24979
UT WOS:000327547200003
PM 23765106
ER

PT J
AU Gold, MG
Fowler, DM
Means, CK
Pawson, CT
Stephany, JJ
Langeberg, LK
Fields, S
Scott, JD
AF Gold, Matthew G.
Fowler, Douglas M.
Means, Christopher K.
Pawson, Catherine T.
Stephany, Jason J.
Langeberg, Lorene K.
Fields, Stanley
Scott, John D.
TI Engineering A-kinase Anchoring Protein (AKAP)-selective Regulatory Subunits of Protein Kinase A (PKA) through Structure-based Phage Selection
SO JOURNAL OF BIOLOGICAL CHEMISTRY
AB PKA is retained within distinct subcellular environments by the association of its regulatory type II (RII) subunits with A-kinase anchoring proteins (AKAPs). Conventional reagents that universally disrupt PKA anchoring are patterned after a conserved AKAP motif. We introduce a phage selection procedure that exploits high-resolution structural information to engineer RII mutants that are selective for a particular AKAP. Selective RII (R-Select) sequences were obtained for eight AKAPs following competitive selection screening. Biochemical and cell-based experiments validated the efficacy of RSelect proteins for AKAP2 and AKAP18. These engineered proteins represent a new class of reagents that can be used to dissect the contributions of different AKAP-targeted pools of PKA. Molecular modeling and high-throughput sequencing analyses revealed the molecular basis of AKAP-selective interactions and shed new light on native RII-AKAP interactions. We propose that this structure-directed evolution strategy might be generally applicable for the investigation of other protein interaction surfaces.
SN 0021-9258
PD JUN 14
PY 2013
VL 288

IS 24
BP 17111
EP 17121
DI 10.1074/jbc.M112.447326
UT WOS:000320380600009
PM 23625929
ER

PT J
AU Jiang, L
Mishra, P
Hietpas, RT
Zeldovich, KB
Bolon, DNA
AF Jiang, Li
Mishra, Parul
Hietpas, Ryan T.
Zeldovich, Konstantin B.
Bolon, Daniel N. A.
TI Latent Effects of Hsp90 Mutants Revealed at Reduced Expression Levels
SO PLOS GENETICS
AB In natural systems, selection acts on both protein sequence and expression level, but it is unclear how selection integrates over these two dimensions. We recently developed the EMPIRIC approach to systematically determine the fitness effects of all possible point mutants for important regions of essential genes in yeast. Here, we systematically investigated the fitness effects of point mutations in a putative substrate binding loop of yeast Hsp90 (Hsp82) over a broad range of expression strengths. Negative epistasis between reduced expression strength and amino acid substitutions was common, and the endogenous expression strength frequently obscured mutant defects. By analyzing fitness effects at varied expression strengths, we were able to uncover all mutant effects on function. The majority of mutants caused partial functional defects, consistent with this region of Hsp90 contributing to a mutation sensitive and critical process. These results demonstrate that important functional regions of proteins can tolerate mutational defects without experimentally observable impacts on fitness.
OI Bolon, Daniel/0000-0001-5857-6676
SN 1553-7404
PD JUN
PY 2013
VL 9
IS 6
AR e1003600
DI 10.1371/journal.pgen.1003600
UT WOS:000321222600066
PM 23825969
ER

PT J
AU McCandlish, DM
Rajon, E
Shah, P
Ding, Y
Plotkin, JB
AF McCandlish, David M.
Rajon, Etienne
Shah, Premal
Ding, Yang
Plotkin, Joshua B.
TI The role of epistasis in protein evolution
SO NATURE
SN 0028-0836
PD MAY 30
PY 2013
VL 497
IS 7451
BP E1
EP E2
DI 10.1038/nature12219
UT WOS:000319556100001
PM 23719465
ER

PT J

AU Roscoe, BP
Thayer, KM
Zeldovich, KB
Fushman, D
Bolon, DNA
AF Roscoe, Benjamin P.
Thayer, Kelly M.
Zeldovich, Konstantin B.
Fushman, David
Bolon, Daniel N. A.
TI Analyses of the Effects of All Ubiquitin Point Mutants on Yeast Growth Rate
SO JOURNAL OF MOLECULAR BIOLOGY
AB The amino acid sequence of a protein governs its function. We used bulk competition and focused deep sequencing to investigate the effects of all ubiquitin point mutants on yeast growth rate. Many aspects of ubiquitin function have been carefully studied, which enabled interpretation of our growth analyses in light of a rich structural, biophysical and biochemical knowledge base. In one highly sensitive cluster on the surface of ubiquitin, almost every amino acid substitution caused growth defects. In contrast, the opposite face tolerated virtually all possible substitutions. Surface locations between these two faces exhibited intermediate mutational tolerance. The sensitive face corresponds to the known interface for many binding partners. Across all surface positions, we observe a strong correlation between burial at structurally characterized interfaces and the number of amino acid substitutions compatible with robust growth. This result indicates that binding is a dominant determinant of ubiquitin function. In the solvent-inaccessible core of ubiquitin, all positions tolerated a limited number of substitutions, with hydrophobic amino acids especially interchangeable. Some mutations null for yeast growth were previously shown to populate folded conformations indicating that, for these mutants, subtle changes to conformation caused functional defects. The most sensitive region to mutation within the core was located near the C-terminus that is a focal binding site for many critical binding partners. These results indicate that core mutations may frequently cause functional defects through subtle disturbances to structure or dynamics. (C) 2013 Elsevier Ltd. All rights reserved.
OI Bolon, Daniel/0000-0001-5857-6676
SN 0022-2836
EI 1089-8638
PD APR 26
PY 2013
VL 425
IS 8
BP 1363
EP 1377
DI 10.1016/j.jmb.2013.01.032
UT WOS:000317796200010
PM 23376099
ER

PT J
AU Starita, LM
Pruneda, JN
Lo, RS
Fowler, DM
Kim, HJ
Hiatt, JB
Shendure, J
Brzovic, PS
Fields, S
Klevit, RE
AF Starita, Lea M.
Pruneda, Jonathan N.
Lo, Russell S.
Fowler, Douglas M.
Kim, Helen J.
Hiatt, Joseph B.
Shendure, Jay
Brzovic, Peter S.
Fields, Stanley
Klevit, Rachel E.
TI Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis
SO PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA
AB Although ubiquitination plays a critical role in virtually all cellular processes, mechanistic details of ubiquitin (Ub) transfer are still being defined. To identify the molecular determinants within E3 ligases that modulate activity, we scored each member of a library of nearly 100,000 protein variants of the murine ubiquitination factor E4B (Ube4b) U-box domain for auto-ubiquitination activity in the presence of the E2 UbcH5c. This assay identified

mutations that enhance activity both in vitro and in cellular p53 degradation assays. The activity-enhancing mutations fall into two distinct mechanistic classes: One increases the U-box: E2-binding affinity, and the other allosterically stimulates the formation of catalytically active conformations of the E2 similar to Ub conjugate. The same mutations enhance E3 activity in the presence of another E2, Ube2w, implying a common allosteric mechanism, and therefore the general applicability of our observations to other E3s. A comparison of the E3 activity with the two different E2s identified an additional variant that exhibits E3:E2 specificity. Our results highlight the general utility of high-throughput mutagenesis in delineating the molecular basis of enzyme activity.
OI Shendure, Jay/0000-0002-1516-1865
SN 0027-8424
PD APR 2
PY 2013
VL 110
IS 14
BP E1263
EP E1272
DI 10.1073/pnas.1303309110
UT WOS:000318037800007
PM 23509263
ER

PT J
AU Podgornaia, AI
Laub, MT
AF Podgornaia, Anna I.
Laub, Michael T.
TI Determinants of specificity in two-component signal transduction
SO CURRENT OPINION IN MICROBIOLOGY
AB Maintaining the faithful flow of information through signal transduction pathways is critical to the survival and proliferation of organisms. This problem is particularly challenging as many signaling proteins are part of large, paralogous families that are highly similar at the sequence and structural levels, increasing the risk of unwanted cross-talk. To detect environmental signals and process information, bacteria rely heavily on two-component signaling systems comprised of sensor histidine kinases and their cognate response regulators. Although most species encode dozens of these signaling pathways, there is relatively little cross-talk, indicating that individual pathways are well insulated and highly specific. Here, we review the molecular mechanisms that enforce this specificity. Further, we highlight recent studies that have revealed how these mechanisms evolve to accommodate the introduction of new pathways by gene duplication.
OI , /0000-0002-8288-7607
SN 1369-5274
PD APR
PY 2013
VL 16
IS 2
BP 156
EP 162
DI 10.1016/j.mib.2013.01.004
UT WOS:000319180100008
PM 23352354
ER

PT J
AU Liachko, I
Youngblood, RA
Keich, U
Dunham, MJ
AF Liachko, Ivan
Youngblood, Rachel A.
Keich, Uri
Dunham, Maitreya J.
TI High-resolution mapping, characterization, and optimization of autonomously replicating sequences in yeast
SO GENOME RESEARCH
AB DNA replication origins are necessary for the duplication of genomes. In addition, plasmid-based expression systems require DNA replication origins to maintain plasmids efficiently. The yeast autonomously replicating sequence (ARS) assay has been a valuable tool in dissecting replication origin structure and function. However, the dearth of information on origins in diverse yeasts limits the availability of efficient replication origin modules to only a handful of species and restricts our understanding of origin function and evolution. To enable rapid study of origins, we have developed a sequencing-based suite of methods for comprehensively mapping and characterizing ARSs within a yeast genome. Our approach finely maps genomic inserts capable of supporting plasmid replication and uses massively parallel deep mutational scanning to define molecular determinants of ARS function with single-nucleotide resolution. In addition to providing unprecedented detail into origin structure, our data have allowed us to design short, synthetic DNA sequences that retain maximal ARS function. These methods can be readily applied to understand and modulate ARS function in diverse systems.
OI Dunham, Maitreya/0000-0001-9944-2666

SN 1088-9051
PD APR
PY 2013
VL 23
IS 4
BP 698
EP 704
DI 10.1101/gr.144659.112
UT WOS:000316920500011
PM 23241746
ER

PT J
AU Bhattacharyya, S
  Varadarajan, R
AF Bhattacharyya, Sanchari
  Varadarajan, Raghavan
TI Packing in molten globules and native states
SO CURRENT OPINION IN STRUCTURAL BIOLOGY
AB Close packing of hydrophobic residues in the protein interior is an important determinant of protein stability. Cavities introduced by large to small substitutions are known to destabilize proteins. Conversely, native states of proteins and protein fragments can be stabilized by filling in existing cavities. Molten globules (MGs) were initially used to describe a state of protein which has well-defined secondary structure but little or no tertiary packing. Subsequent studies have shown that MGs do have some degree of native-like topology and specific packing. Wet molten globules (WMGs) with hydrated cores and considerably decreased packing relative to the native state have been studied extensively. Recently there has been renewed interest in identification and characterization of dry molten globules (DMGs). These are slightly expanded forms of the native state which show increased conformational flexibility, native-like main-chain hydrogen bonding and dry interiors. The generality of occurrence of DMGs during protein unfolding and the extent and nature of packing in DMGs remain to be elucidated. Packing interactions in native proteins and MGs can be probed through mutations. Next generation sequencing technologies make it possible to determine relative populations of mutants in a large pool. When this is coupled to phenotypic screens or cell-surface display, it becomes possible to rapidly examine large panels of single-site or multi-site mutants. From such studies, residue specific contributions to protein stability and function can be estimated in a highly parallelized fashion. This complements conventional biophysical methods for characterization of packing in native states and molten globules.
SN 0959-440X
PD FEB
PY 2013
VL 23
IS 1
BP 11
EP 21
DI 10.1016/j.sbi.2012.10.010
UT WOS:000315832700003
PM 23270864
ER

PT J
AU Studer, RA
  Dessailly, BH
  Orengo, CA
AF Studer, Romain A.
  Dessailly, Benoit H.
  Orengo, Christine A.
TI Residue mutations and their impact on protein structure and function:
  detecting beneficial and pathogenic changes
SO BIOCHEMICAL JOURNAL
AB The present review focuses on the evolution of proteins and the impact of amino acid mutations on function from a structural perspective. Proteins evolve under the law of natural selection and undergo alternating periods of conservative evolution and of relatively rapid change. The likelihood of mutations being fixed in the genome depends on various factors, such as the fitness of the phenotype or the position of the residues in the three-dimensional structure. For example, co-evolution of residues located close together in three-dimensional space can occur to preserve global stability. Whereas point mutations can fine-tune the protein function, residue insertions and deletions ('decorations' at the structural level) can sometimes modify functional sites and protein interactions more dramatically. We discuss recent developments and tools to identify such episodic mutations, and examine their applications in medical research. Such tools have been tested on simulated data and applied to real data such as viruses or animal sequences. Traditionally, there has been little if any crosstalk between the fields of protein biophysics, protein structure function and molecular evolution. However, the last several years have seen some exciting developments in combining these approaches to obtain an in-depth understanding of how proteins evolve. For example, a better understanding of how structural constraints affect protein evolution will greatly help us to optimize our models of sequence evolution. The present review explores this new synthesis of perspectives.
RI Studer, Romain/F-8141-2011
OI Studer, Romain/0000-0003-0687-9848
SN 0264-6021

EI 1470-8728
PD FEB 1
PY 2013
VL 449
BP 581
EP 594
DI 10.1042/BJ20121221
PN 3
UT WOS:000313776000002
PM 23301657
ER

PT J
AU Deriziotis, P
  Fisher, SE
AF Deriziotis, Pelagia
  Fisher, Simon E.
TI Neurogenomics of speech and language disorders: the road ahead
SO GENOME BIOLOGY
AB Next-generation sequencing is set to transform the discovery of genes underlying neurodevelopmental disorders, and so off er important in sights into the biological bases of spoken language. Success will depend on functional assessments in neuronal cell lines, animal models and humans themselves.
RI Fisher, Simon/E-9130-2012; Derizioti, Pelagia/C-3857-2015
OI Fisher, Simon/0000-0002-3132-1996; Derizioti,
  Pelagia/0000-0001-5544-8345
SN 1465-6906
PY 2013
VL 14
IS 4
AR 204
DI 10.1186/gb-2013-14-4-204
UT WOS:000322521300015
PM 23597266
ER

PT S
AU Whitehead, TA
  Baker, D
  Fleishman, SJ
AF Whitehead, Timothy A.
  Baker, David
  Fleishman, Sarel J.
BE Keating, AE
TI Computational Design of Novel Protein Binders and Experimental Affinity
  Maturation
SO METHODS IN PROTEIN DESIGN
SE Methods in Enzymology
AB Computational design of novel protein binders has recently emerged as a useful technique to study biomolecular recognition and generate molecules for use in biotechnology, research, and biomedicine. Current limitations in computational design methodology have led to the adoption of high-throughput screening and affinity maturation techniques to diagnose modeling inaccuracies and generate high activity binders. Here, we scrutinize this combination of computational and experimental aspects and propose areas for future methodological improvements.
RI Baker, David/K-8941-2012;
OI Baker, David/0000-0001-7896-6217; Fleishman, Sarel/0000-0003-3177-7560
SN 0076-6879
BN 978-0-12-394292-0
PY 2013
VL 523
BP 1
EP 19
DI 10.1016/B978-0-12-394292-0.00001-1
UT WOS:000318253400002
PM 23422423
ER

PT S
AU Ashenberg, O
  Laub, MT

AF Ashenberg, Orr
Laub, Michael T.
BE Keating, AE
TI Using Analyses of Amino Acid Coevolution to Understand Protein Structure and Function
SO METHODS IN PROTEIN DESIGN
SE Methods in Enzymology
AB Determining which residues of a protein contribute to a specific function is a difficult problem. Analyses of amino acid covariation within a protein family can serve as a useful guide by identifying residues that are functionally coupled. Covariation analyses have been successfully used on several different protein families to identify residues that work together to promote folding, enable protein protein interactions, or contribute to an enzymatic activity. Covariation is a statistical signal that can be measured in a multiple sequence alignment of homologous proteins. As sequence databases have expanded dramatically, covariation analyses have become easier and more powerful. In this chapter, we describe how functional covariation arises during the evolution of proteins and how this signal can be distinguished from various background signals. We discuss the basic methodology for performing amino acid covariation analysis, using bacterial two-component signal transduction proteins as an example. We provide practical suggestions for each step of the process including assembly of protein sequences, construction of a multiple sequence alignment measurement of covariation, and analysis of results.
SN 0076-6879
BN 978-0-12-394292-0
PY 2013
VL 523
BP 191
EP 212
DI 10.1016/B978-0-12-394292-0.00009-6
UT WOS:000318253400010
PM 23422431
ER

PT J
AU Wu, NC
Young, AP
Dandekar, S
Wijersuriya, H
Al-Mawsawi, LQ
Wu, TT
Sun, R
AF Wu, Nicholas C.
Young, Arthur P.
Dandekar, Sugandha
Wijersuriya, Hemani
Al-Mawsawi, Laith Q.
Wu, Ting-Ting
Sun, Ren
TI Systematic Identification of H274Y Compensatory Mutations in Influenza A Virus Neuraminidase by High-Throughput Screening
SO JOURNAL OF VIROLOGY
AB Compensatory mutations contribute to the appearance of the oseltamivir resistance substitution H274Y in the neuraminidase (NA) gene of H1N1 influenza viruses. Here, we describe a high-throughput screening method utilizing error-prone PCR and next-generation sequencing to comprehensively screen NA genes for H274Y compensatory mutations. We found four mutations that can either fully (R194G, E214D) or partially (L250P, F239Y) compensate for the fitness deficiency of the H274Y mutant. The compensatory effect of E214D is applicable in both seasonal influenza virus strain A/New Caledonia/20/1999 and 2009 pandemic swine influenza virus strain A/California/04/2009. The technique described here has the potential to profile a gene at the single-nucleotide level to comprehend the dynamics of mutation space and fitness and thus offers prediction power for emerging mutant species.
RI Wu, Nicholas/H-3822-2015
OI Wu, Nicholas/0000-0002-9078-6697
SN 0022-538X
PD JAN
PY 2013
VL 87
IS 2
BP 1193
EP 1199
DI 10.1128/JVI.01658-12
UT WOS:000312934400045
PM 23152521
ER

PT B

AU Stapleton, JA
Rodriguez-Granillo, A
Nanda, V
AF Stapleton, James A.
Rodriguez-Granillo, Agustina
Nanda, Vikas
BE Xie, Y
TI Artificial Enzymes
SO NANOBIOTECHNOLOGY HANDBOOK
BN 978-1-4398-3870-9; 978-1-4398-3869-3
PY 2013
BP 47
EP 71
UT WOS:000355554400004
ER

PT J
AU Deng, ZF
Huang, WZ
Bakkalbasi, E
Brown, NG
Adamski, CJ
Rice, K
Muzny, D
Gibbs, RA
Palzkill, T
AF Deng, Zhifeng
Huang, Wanzhi
Bakkalbasi, Erol
Brown, Nicholas G.
Adamski, Carolyn J.
Rice, Kacie
Muzny, Donna
Gibbs, Richard A.
Palzkill, Timothy
TI Deep Sequencing of Systematic Combinatorial Libraries Reveals
beta-Lactamase Sequence Constraints at High Resolution
SO JOURNAL OF MOLECULAR BIOLOGY
AB In this study, combinatorial libraries were used in conjunction with ultrahigh-throughput sequencing to comprehensively determine the impact of each of the 19 possible amino acid substitutions at each residue position in the TEM-1 beta-lactamase enzyme. The libraries were introduced into Escherichia coli, and mutants were selected for ampicillin resistance. The selected colonies were pooled and subjected to ultrahigh-throughput sequencing to reveal the sequence preferences at each position. The depth of sequencing provided a clear, statistically significant picture of what amino acids are favored for ampicillin hydrolysis for all 263 positions of the enzyme in one experiment. Although the enzyme is generally tolerant of amino acid substitutions, several surface positions far from the active site are sensitive to substitutions suggesting a role for these residues in enzyme stability, solubility, or catalysis. In addition, information on the frequency of substitutions was used to identify mutations that increase enzyme thermodynamic stability. Finally, a comparison of sequence requirements based on the mutagenesis results versus those inferred from sequence conservation in an alignment of 156 class A beta-lactamases reveals significant differences in that several residues in TEM-1 do not tolerate substitutions and yet extensive variation is observed in the alignment and vice versa. An analysis of the TEM-1 and other class A structures suggests that residues that vary in the "alignment may nevertheless make unique, but important, interactions within individual enzymes. (C) 2012 Elsevier Ltd. All rights reserved.
SN 0022-2836
EI 1089-8638
PD DEC 7
PY 2012
VL 424
IS 3-4
BP 150
EP 167
DI 10.1016/j.jmb.2012.09.014
UT WOS:000312753700005
PM 23017428
ER

PT J
AU Fujino, Y
Fujita, R
Wada, K
Fujishige, K

Kanamori, T
Hunt, L
Shimizu, Y
Ueda, T
AF Fujino, Yasuhiro
Fujita, Risako
Wada, Kouichi
Fujishige, Kotomi
Kanamori, Takashi
Hunt, Lindsey
Shimizu, Yoshihiro
Ueda, Takuya
TI Robust in vitro affinity maturation strategy based on interface-focused high-throughput mutational scanning
SO BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS
AB Development of protein therapeutics or biosensors often requires in vitro affinity maturation. Here we report a robust affinity engineering strategy using a custom designed library. The strategy consists of two steps beginning with identification of beneficial single amino acid substitutions then combination. A high quality combinatorial library specifically customized to a given binding-interface can be rapidly designed by high-throughput mutational scanning of single substitution libraries. When applied to the optimization of a model antibody Fab fragment, the strategy created a diverse panel of high affinity variants. The most potent variant achieved 2110-fold affinity improvement to an equilibrium dissociation constant (Kd) of 3.45 pM with only 7 amino acid substitutions. The method should facilitate affinity engineering of a wide variety of protein-protein interactions due to its context-dependent library design strategy. (C) 2012 Elsevier Inc. All rights reserved.
RI Shimizu, Yoshihiro/A-6472-2016; Ueda, Takuya/K-5217-2014
OI Ueda, Takuya/0000-0002-7760-8271
SN 0006-291X
PD NOV 23
PY 2012
VL 428
IS 3
BP 395
EP 400
DI 10.1016/j.bbrc.2012.10.066
UT WOS:000313021700013
PM 23103372
ER

PT J
AU McLaughlin, RN
Poelwijk, FJ
Raman, A
Gosal, WS
Ranganathan, R
AF McLaughlin, Richard N., Jr.
Poelwijk, Frank J.
Raman, Arjun
Gosal, Walraj S.
Ranganathan, Rama
TI The spatial architecture of protein function and adaptation
SO NATURE
AB Statistical analysis of protein evolution suggests a design for natural proteins in which sparse networks of coevolving amino acids (termed sectors) comprise the essence of three-dimensional structure and function(1-5). However, proteins are also subject to pressures deriving from the dynamics of the evolutionary process itself-the ability to tolerate mutation and to be adaptive to changing selection pressures(6-10). To understand the relationship of the sector architecture to these properties, we developed a high-throughput quantitative method for a comprehensive single-mutation study in which every position is substituted individually to every other amino acid. Using a PDZ domain (PSD95(pdz3)) model system, we show that sector positions are functionally sensitive to mutation, whereas non-sector positions are more tolerant to substitution. In addition, we find that adaptation to a new binding specificity initiates exclusively through variation within sector residues. A combination of just two sector mutations located near and away from the ligand-binding site suffices to switch the binding specificity of PSD95(pdz3) quantitatively towards a class-switching ligand. The localization of functional constraint and adaptive variation within the sector has important implications for understanding and engineering proteins.
RI Gosal, Walraj/J-5260-2012;
OI Gosal, Walraj/0000-0003-3396-1505; Poelwijk, Frank/0000-0002-5696-4357
SN 0028-0836
PD NOV 1
PY 2012
VL 491
IS 7422
BP 138
EP U163

DI 10.1038/nature11500
UT WOS:000310434500046
PM 23041932
ER

PT J
AU Dolled-Filhart, MP
Lordemann, A
Dahl, W
Haraksingh, RR
Ou-Yang, CW
Lin, JCH
AF Dolled-Filhart, Marisa P.
Lordemann, Amanda
Dahl, William
Haraksingh, Rajini Rani
Ou-Yang, Chih-Wen
Lin, Jimmy Cheng-Ho
TI Personalizing rare disease research: how genomics is revolutionizing the diagnosis and treatment of rare disease
SO PERSONALIZED MEDICINE
AB A decade after the complete sequencing of the human genome, combined with recent advances in throughput and sequencing costs, the genetics of rare diseases has entered a new era. There has now been an explosion in the identification and mapping of rare diseases, with over 10,000 exomes having been sequenced to date. This article surveys the progress and development of technologies to understand rare disease; it provides a historical overview of traditional techniques such as karyotyping and homozygosity mapping, reviews current methods of whole-exome and -genome sequencing, and provides a future perspective on upcoming developments such as targeted drugs and gene therapy. This article will discuss the implications of these methods for rare disease research, along with a discussion of the success stories that provide great hope and optimism for patients and scientists alike.
OI Haraksingh, Rajini/0000-0002-6644-8874
SN 1741-0541
PD NOV
PY 2012
VL 9
IS 8
BP 805
EP 819
DI 10.2217/PME.12.97
UT WOS:000311977800007
ER

PT J
AU Shendure, J
Aiden, EL
AF Shendure, Jay
Aiden, Erez Lieberman
TI The expanding scope of DNA sequencing
SO NATURE BIOTECHNOLOGY
AB In just seven years, next-generation technologies have reduced the cost and increased the speed of DNA sequencing by four orders of magnitude, and experiments requiring many millions of sequencing reads are now routine. In research, sequencing is being applied not only to assemble genomes and to investigate the genetic basis of human disease, but also to explore myriad phenomena in organismic and cellular biology. In the clinic, the utility of sequence data is being intensively evaluated in diverse contexts, including reproductive medicine, oncology and infectious disease. A recurrent theme in the development of new sequencing applications is the creative 'recombination' of existing experimental building blocks. However, there remain many potentially high-impact applications of next-generation DNA sequencing that are not yet fully realized.
OI Shendure, Jay/0000-0002-1516-1865
SN 1087-0156
EI 1546-1696
PD NOV
PY 2012
VL 30
IS 11
BP 1084
EP 1094
DI 10.1038/nbt.2421
UT WOS:000311087500028
PM 23138308
ER

PT J

AU Xu, ZH
Juan, V
Ivanov, A
Ma, ZY
Polakoff, D
Powers, DB
DuBridge, RB
Wilson, K
Akamatsu, Y
AF Xu, Zhenghai
Juan, Veronica
Ivanov, Alexander
Ma, Zhiyuan
Polakoff, Dixie
Powers, David B.
DuBridge, Robert B.
Wilson, Keith
Akamatsu, Yoshiko
TI Affinity and Cross-Reactivity Engineering of CTLA4-Ig To Modulate T Cell
Costimulation
SO JOURNAL OF IMMUNOLOGY
AB CTLA4-Ig is an Fc fusion protein containing the extracellular domain of CTLA-4, a receptor known to deliver a negative signal to T cells. CTLA4-Ig modulates T cell costimulatory signals by blocking the CD80 and CD86 ligands from binding to CD28, which delivers a positive T cell costimulatory signal. To engineer CTLA4-Ig variants with altered binding affinity to CD80 and CD86, we employed a high-throughput protein engineering method to map the ligand binding surface of CTLA-4. The resulting mutagenesis map identified positions critical for the recognition of each ligand on the three CDR-like loops of CTLA-4, consistent with the published site-directed mutagenesis and x-ray crystal structures of the CTLA-4/CD80 and CTLA-4/CD86 complexes. A number of single amino acid substitutions were identified that equally affected the binding affinity of CTLA4-Ig for both ligands as well as those that differentially affected binding. All of the high-affinity variants showed improved off-rates, with the best one being a 17.5-fold improved off-rate over parental CTLA4-Ig binding to CD86. Allostimulation of human CD4(+) T cells showed that improvement of CD80 and CD86 binding activity augmented inhibition of naive and primed T cell activation. In general, increased affinity for CD86 resulted in more potent inhibition of T cell response than did increased affinity for CD80. Optimization of the affinity balance to CD80 and CD86 to particular disease settings may lead to development of a CTLA4-Ig molecule with improved efficacy and safety profiles. The Journal of Immunology, 2012, 189: 4470-4477.
SN 0022-1767
PD NOV 1
PY 2012
VL 189
IS 9
BP 4470
EP 4477
DI 10.4049/jimmunol.1201813
UT WOS:000310200600034
PM 23018459
ER

PT J
AU Traxlmayr, MW
Hasenhindl, C
Hackl, M
Stadlmayr, G
Rybka, JD
Borth, N
Grillari, J
Ruker, F
Obinger, C
AF Traxlmayr, Michael W.
Hasenhindl, Christoph
Hackl, Matthias
Stadlmayr, Gerhard
Rybka, Jakub D.
Borth, Nicole
Grillari, Johannes
Rueker, Florian
Obinger, Christian
TI Construction of a Stability Landscape of the CH3 Domain of Human IgG1 by
Combining Directed Evolution with High Throughput Sequencing
SO JOURNAL OF MOLECULAR BIOLOGY
AB One of the most important but still poorly understood issues in protein chemistry is the relationship between sequence and stability of proteins. Here,

we present a method for analyzing the influence of each individual residue on the foldability and stability of an entire protein. A randomly mutated library of the crystallizable fragment of human immunoglobulin G class 1 (IgG1-Fc) was expressed on the surface of yeast, followed by heat incubation at 79 degrees C and selection of stable variants that still bound to structurally specific ligands. High throughput sequencing allowed comparison of the mutation rate between the starting and selected library pools, enabling the generation of a stability landscape for the entire CH3 domain of human IgG1 at single residue resolution. Its quality was analyzed with respect to (i) the structure of IgG1-Fc, (ii) evolutionarily conserved positions and (iii) in silico calculations of the energy of unfolding of all variants in comparison with the wild-type protein. In addition, this new experimental approach allowed the assignment of functional epitopes of structurally specific ligands used for selection [Fc gamma-receptor I (CD64) and anti-human CH2 domain antibody] to distinct binding regions in the CH2 domain. (C) 2012 Elsevier Ltd. All rights reserved.
RI Grillari, Johannes/B-2967-2011
OI Obinger, Christian/0000-0002-7133-3430; Grillari,
Johannes/0000-0001-5474-6332
SN 0022-2836
PD OCT 26
PY 2012
VL 423
IS 3
BP 397
EP 412
DI 10.1016/j.jmb.2012.07.017
UT WOS:000310415400010
PM 22846908
ER

PT J
AU Araya, CL
Fowler, DM
Chen, WT
Muniez, I
Kelly, JW
Fields, S
AF Araya, Carlos L.
Fowler, Douglas M.
Chen, Wentao
Muniez, Ike
Kelly, Jeffery W.
Fields, Stanley
TI A fundamental protein property, thermodynamic stability, revealed solely
from large-scale measurements of protein function
SO PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF
AMERICA
AB The ability of a protein to carry out a given function results from fundamental physicochemical properties that include the protein's structure, mechanism of action, and thermodynamic stability. Traditional approaches to study these properties have typically required the direct measurement of the property of interest, oftentimes a laborious undertaking. Although protein properties can be probed by mutagenesis, this approach has been limited by its low through-put. Recent technological developments have enabled the rapid quantification of a protein's function, such as binding to a ligand, for numerous variants of that protein. Here, we measure the ability of 47,000 variants of a WW domain to bind to a peptide ligand and use these functional measurements to identify stabilizing mutations without directly assaying stability. Our approach is rooted in the well-established concept that protein function is closely related to stability. Protein function is generally reduced by destabilizing mutations, but this decrease can be rescued by stabilizing mutations. Based on this observation, we introduce partner potentiation, a metric that uses this rescue ability to identify stabilizing mutations, and identify 15 candidate stabilizing mutations in the WW domain. We tested six candidates by thermal denaturation and found two highly stabilizing mutations, one more stabilizing than any previously known mutation. Thus, physicochemical properties such as stability are latent within these large-scale protein functional data and can be revealed by systematic analysis. This approach should allow other protein properties to be discovered.
OI Araya, Carlos/0000-0002-5512-3062
SN 0027-8424
PD OCT 16
PY 2012
VL 109
IS 42
BP 16858
EP 16863
DI 10.1073/pnas.1209751109
UT WOS:000310515800030
PM 23035249
ER

PT J
AU Moal, IH
Fernandez-Recio, J

AF Moal, Iain H.
Fernandez-Recio, Juan
TI SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models
SO BIOINFORMATICS
AB Motivation: Empirical models for the prediction of how changes in sequence alter protein-protein binding kinetics and thermodynamics can garner insights into many aspects of molecular biology. However, such models require empirical training data and proper validation before they can be widely applied. Previous databases contained few stabilizing mutations and no discussion of their inherent biases or how this impacts model construction or validation.
Results: We present SKEMPI, a database of 3047 binding free energy changes upon mutation assembled from the scientific literature, for protein-protein heterodimeric complexes with experimentally determined structures. This represents over four times more data than previously collected. Changes in 713 association and dissociation rates and 127 enthalpies and entropies were also recorded. The existence of biases towards specific mutations, residues, interfaces, proteins and protein families is discussed in the context of how the data can be used to construct predictive models. Finally, a cross-validation scheme is presented which is capable of estimating the efficacy of derived models on future data in which these biases are not present.
OI Fernandez-Recio, Juan/0000-0002-3986-7686; Moal, Iain/0000-0002-4960-5487
SN 1367-4803
PD OCT 15
PY 2012
VL 28
IS 20
BP 2600
EP 2607
DI 10.1093/bioinformatics/bts489
UT WOS:000309881200006
PM 22859501
ER

PT J
AU Uyttendaele, I
Lavens, D
Catteeuw, D
Lemmens, I
Bovijn, C
Tavernier, J
Peelman, F
AF Uyttendaele, Isabel
Lavens, Delphine
Catteeuw, Dominiek
Lemmens, Irma
Bovijn, Celia
Tavernier, Jan
Peelman, Frank
TI Random Mutagenesis MAPPIT Analysis Identifies Binding Sites for Vif and Gag in Both Cytidine Deaminase Domains of Apobec3G
SO PLOS ONE
AB The mammalian two-hybrid system MAPPIT allows the detection of protein-protein interactions in intact human cells. We developed a random mutagenesis screening strategy based on MAPPIT to detect mutations that disrupt the interaction of one protein with multiple protein interactors simultanously. The strategy was used to detect residues of the human cytidine deaminase Apobec3G that are important for its homodimerization and its interaction with the HIV-1 Gag and Vif proteins. The strategy is able to identify the previously described head-to-head homodimerization interface in the N-terminal domain of Apobec3G. Our analysis further detects two new potential interaction surfaces in the N- and C-terminal domain of Apobec3G for interaction with Vif and Gag or for Apobec3G dimerization.
SN 1932-6203
PD SEP 10
PY 2012
VL 7
IS 9
AR e44143
DI 10.1371/journal.pone.0044143
UT WOS:000308748400011
PM 22970171
ER

PT J
AU Matochko, WL
Chu, KK

Jin, BJ
Lee, SW
Whitesides, GM
Derda, R
AF Matochko, Wadim L.
Chu, Kiki
Jin, Bingjie
Lee, Sam W.
Whitesides, George M.
Derda, Ratmir
TI Deep sequencing analysis of phage libraries using Illumina platform
SO METHODS
AB This paper presents an analysis of phage-displayed libraries of peptides using Illumina. We describe steps for the preparation of short DNA fragments for deep sequencing and MatLab software for the analysis of the results. Screening of peptide libraries displayed on the surface of bacteriophage (phage display) can be used to discover peptides that bind to any target. The key step in this discovery is the analysis of peptide sequences present in the library. This analysis is usually performed by Sanger sequencing, which is labor intensive and limited to examination of a few hundred phage clones. On the other hand. Illumina deep-sequencing technology can characterize over 10(7) reads in a single run. We applied Illumina sequencing to analyze phage libraries. Using PCR, we isolated the variable regions from M13KE phage vectors from a phage display library. The PCR primers contained (i) sequences flanking the variable region, (ii) barcodes, and (iii) variable 5'-terminal region. We used this approach to examine how diversity of peptides in phage display libraries changes as a result of amplification of libraries in bacteria. Using HiSeq single-end Illumina sequencing of these fragments, we acquired over 2 x 10(7) reads, 57 base pairs (bp) in length. Each read contained information about the barcode (6 bp), one complimentary region (12 bp) and a variable region (36 bp). We applied this sequencing to a model library of 10(6) unique clones and observed that amplification enriches similar to 150 clones, which dominate similar to 20% of the library. Deep sequencing, for the first time, characterized the collapse of diversity in phage libraries. The results suggest that screens based on repeated amplification and small-scale sequencing identify a few binding clones and miss thousands of useful clones. The deep sequencing approach described here could identify under-represented clones in phage screens. It could also be instrumental in developing new screening strategies, which can preserve diversity of phage clones and identify ligands previously lost in phage display screens. (C) 2012 Elsevier Inc. All rights reserved.
SN 1046-2023
PD SEP
PY 2012
VL 58
IS 1
BP 47
EP 55
DI 10.1016/j.ymeth.2012.07.006
UT WOS:000311526000008
PM 22819855
ER

PT J
AU Creixell, P
Schoof, EM
Erler, JT
Linding, R
AF Creixell, Pau
Schoof, Erwin M.
Erler, Janine T.
Linding, Rune
TI Navigating cancer network attractors for tumor-specific therapy
SO NATURE BIOTECHNOLOGY
AB Cells employ highly dynamic signaling networks to drive biological decision processes. Perturbations to these signaling networks may attract cells to new malignant signaling and phenotypic states, termed cancer network attractors, that result in cancer development. As different cancer cells reach these malignant states by accumulating different molecular alterations, uncovering these mechanisms represents a grand challenge in cancer biology. Addressing this challenge will require new systems-based strategies that capture the intrinsic properties of cancer signaling networks and provide deeper understanding of the processes by which genetic lesions perturb these networks and lead to disease phenotypes. Network biology will help circumvent fundamental obstacles in cancer treatment, such as drug resistance and metastasis, empowering personalized and tumor-specific cancer therapies.
OI Erler, Janine/0000-0001-8675-6527
SN 1087-0156
PD SEP
PY 2012
VL 30
IS 9
BP 842
EP 848
DI 10.1038/nbt.2345
UT WOS:000308705700018
PM 22965061

ER

PT J
AU Teyra, J
Sidhu, SS
Kim, PM
AF Teyra, Joan
Sidhu, Sachdev S.
Kim, Philip M.
TI Elucidation of the binding preferences of peptide recognition modules:
SH3 and PDZ domains
SO FEBS LETTERS
AB Peptide-binding domains play a critical role in regulation of cellular processes by mediating protein interactions involved in signalling. In recent years, the development of large-scale technologies has enabled exhaustive studies on the peptide recognition preferences for a number of peptide-binding domain families. These efforts have provided significant insights into the binding specificities of these modular domains. Many research groups have taken advantage of this unprecedented volume of specificity data and have developed a variety of new algorithms for the prediction of binding specificities of peptide-binding domains and for the prediction of their natural binding targets. This knowledge has also been applied to the design of synthetic peptide-binding domains in order to rewire protein-protein interaction networks. Here, we describe how these experimental technologies have impacted on our understanding of peptide-binding domain specificities and on the elucidation of their natural ligands. We discuss SH3 and PDZ domains as well characterized examples, and we explore the feasibility of expanding high-throughput experiments to other peptide-binding domains. (C) 2012 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
SN 0014-5793
PD AUG 14
PY 2012
VL 586
IS 17
BP 2631
EP 2637
DI 10.1016/j.febslet.2012.05.043
UT WOS:000307295200011
PM 22691579
ER

PT J
AU Gfeller, D
AF Gfeller, David
TI Uncovering new aspects of protein interactions through analysis of
specificity landscapes in peptide recognition domains
SO FEBS LETTERS
AB Protein interactions underlie all biological processes. An important class of protein interactions, often observed in signaling pathways, consists of peptide recognition domains binding short protein segments on the surface of their target proteins. Recent developments in experimental techniques have uncovered many such interactions and shed new lights on their specificity. To analyze these data, novel computational methods have been introduced that can accurately describe the specificity landscape of peptide recognition domains and predict new interactions. Combining large-scale analysis of binding specificity data with structure-based modeling can further reveal new biological insights into the molecular recognition events underlying signaling pathways. (C) 2012 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
SN 0014-5793
EI 1873-3468
PD AUG 14
PY 2012
VL 586
IS 17
BP 2764
EP 2772
DI 10.1016/j.febslet.2012.03.054
UT WOS:000307295200024
PM 22710167
ER

PT J
AU Ryvkin, A
Ashkenazy, H
Smelyanski, L
Kaplan, G
Penn, O
Weiss-Ottolenghi, Y
Privman, E
Ngam, PB

Woodward, JE
May, GD
Bell, C
Pupko, T
Gershoni, JM
AF Ryvkin, Arie
Ashkenazy, Haim
Smelyanski, Larisa
Kaplan, Gilad
Penn, Osnat
Weiss-Ottolenghi, Yael
Privman, Eyal
Ngam, Peter B.
Woodward, James E.
May, Gregory D.
Bell, Callum
Pupko, Tal
Gershoni, Jonathan M.
TI Deep Panning: Steps towards Probing the IgOme
SO PLOS ONE
AB Background: Polyclonal serum consists of vast collections of antibodies, products of differentiated B-cells. The spectrum of antibody specificities is dynamic and varies with age, physiology, and exposure to pathological insults. The complete repertoire of antibody specificities in blood, the IgOme, is therefore an extraordinarily rich source of information-a molecular record of previous encounters as well as a status report of current immune activity. The ability to profile antibody specificities of polyclonal serum at exceptionally high resolution has been an important and serious challenge which can now be overcome.

Methodology/Principal Findings: Here we illustrate the application of Deep Panning, a method that combines the flexibility of combinatorial phage display of random peptides with the power of high-throughput deep sequencing. Deep Panning is first applied to evaluate the quality and diversity of naive random peptide libraries. The production of very large data sets, hundreds of thousands of peptides, has revealed unexpected properties of combinatorial random peptide libraries and indicates correctives to ensure the quality of the libraries generated. Next, Deep Panning is used to analyze a model monoclonal antibody in addition to allowing one to follow the dynamics of biopanning and peptide selection. Finally Deep Panning is applied to profile polyclonal sera derived from HIV infected individuals.

Conclusions/Significance: The ability to generate and characterize hundreds of thousands of affinity-selected peptides creates an effective means towards the interrogation of the IgOme and understanding of the humoral response to disease. Deep Panning should open the door to new possibilities for serological diagnostics, vaccine design and the discovery of the correlates of immunity to emerging infectious agents.
SN 1932-6203
PD AUG 1
PY 2012
VL 7
IS 8
AR e41469
DI 10.1371/journal.pone.0041469
UT WOS:000307212800022
PM 22870226
ER

PT J
AU Hietpas, R
Roscoe, B
Jiang, L
Bolon, DNA
AF Hietpas, Ryan
Roscoe, Benjamin
Jiang, Li
Bolon, Daniel N. A.
TI Fitness analyses of all possible point mutations for regions of genes in yeast
SO NATURE PROTOCOLS
AB Deep sequencing can accurately measure the relative abundance of hundreds of mutations in a single bulk competition experiment, which can give a direct readout of the fitness of each mutant. Here we describe a protocol that we previously developed and optimized to measure the fitness effects of all possible individual codon substitutions for 10-aa regions of essential genes in yeast. Starting with a conditional strain (i.e., a temperature-sensitive strain), we describe how to efficiently generate plasmid libraries of point mutants that can then be transformed to generate libraries of yeast. The yeast libraries are competed under conditions that select for mutant function. Deep-sequencing analyses are used to determine the relative fitness of all mutants. This approach is faster and cheaper per mutant compared with analyzing individually isolated mutants. The protocol can be performed in similar to 4 weeks and many 10-aa regions can be analyzed in parallel.
OI Bolon, Daniel/0000-0001-5857-6676
SN 1754-2189
EI 1750-2799

PD JUL
PY 2012
VL 7
IS 7
BP 1382
EP 1396
DI 10.1038/nprot.2012.069
UT WOS:000305960400010
PM 22722372
ER

PT J
AU Wodak, SJ
AF Wodak, Shoshana J.
TI Next-generation protein engineering targets influenza
SO NATURE BIOTECHNOLOGY
SN 1087-0156
PD JUN
PY 2012
VL 30
IS 6
BP 502
EP 504
DI 10.1038/nbt.2268
UT WOS:000305158600017
PM 22678386
ER

PT J
AU Whitehead, TA
Chevalier, A
Song, YF
Dreyfus, C
Fleishman, SJ
De Mattos, C
Myers, CA
Kamisetty, H
Blair, P
Wilson, IA
Baker, D
AF Whitehead, Timothy A.
Chevalier, Aaron
Song, Yifan
Dreyfus, Cyrille
Fleishman, Sarel J.
De Mattos, Cecilia
Myers, Chris A.
Kamisetty, Hetunandan
Blair, Patrick
Wilson, Ian A.
Baker, David
TI Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing
SO NATURE BIOTECHNOLOGY
AB We show that comprehensive sequence-function maps obtained by deep sequencing can be used to reprogram interaction specificity and to leapfrog over bottlenecks in affinity maturation by combining many individually small contributions not detectable in conventional approaches. We use this approach to optimize two computationally designed inhibitors against H1N1 influenza hemagglutinin and, in both cases, obtain variants with subnanomolar binding affinity. The most potent of these, a 51-residue protein, is broadly cross-reactive against all influenza group 1 hemagglutinins, including human H2, and neutralizes H1N1 viruses with a potency that rivals that of several human monoclonal antibodies, demonstrating that computational design followed by comprehensive energy landscape mapping can generate proteins with potential therapeutic utility.
RI Valle, Ruben/A-7512-2013; Baker, David/K-8941-2012;
OI Baker, David/0000-0001-7896-6217; Fleishman, Sarel/0000-0003-3177-7560
SN 1087-0156
PD JUN
PY 2012
VL 30
IS 6

BP 543
EP +
DI 10.1038/nbt.2214
UT WOS:000305158600024
PM 22634563
ER

PT J
AU Bornscheuer, UT
Huisman, GW
Kazlauskas, RJ
Lutz, S
Moore, JC
Robins, K
AF Bornscheuer, U. T.
Huisman, G. W.
Kazlauskas, R. J.
Lutz, S.
Moore, J. C.
Robins, K.
TI Engineering the third wave of biocatalysis
SO NATURE
AB Over the past ten years, scientific and technological advances have established biocatalysis as a practical and environmentally friendly alternative to traditional metallo- and organocatalysis in chemical synthesis, both in the laboratory and on an industrial scale. Key advances in DNA sequencing and gene synthesis are at the base of tremendous progress in tailoring biocatalysts by protein engineering and design, and the ability to reorganize enzymes into new biosynthetic pathways. To highlight these achievements, here we discuss applications of protein-engineered biocatalysts ranging from commodity chemicals to advanced pharmaceutical intermediates that use enzyme catalysis as a key step.
RI Bornscheuer, Uwe/C-4612-2012; Lutz, Stefan /H-7853-2013
OI Bornscheuer, Uwe/0000-0003-0685-2696;
SN 0028-0836
PD MAY 10
PY 2012
VL 485
IS 7397
BP 185
EP 194
DI 10.1038/nature11117
UT WOS:000303799800031
PM 22575958
ER

PT J
AU Liu, BA
Engelmann, BW
Nash, PD
AF Liu, Bernard A.
Engelmann, Brett W.
Nash, Piers D.
TI High-throughput analysis of peptide-binding modules
SO PROTEOMICS
AB Modular protein interaction domains (PIDs) that recognize linear peptide motifs are found in hundreds of proteins within the human genome. Some PIDs such as SH2, 143-3, Chromo, and Bromo domains serve to recognize posttranslational modification (PTM) of amino acids (such as phosphorylation, acetylation, methylation, etc.) and translate these into discrete cellular responses. Other modules such as SH3 and PSD-95/Discs-large/ZO-1 (PDZ) domains recognize linear peptide epitopes and serve to organize protein complexes based on localization and regions of elevated concentration. In both cases, the ability to nucleate-specific signaling complexes is in large part dependent on the selectivity of a given protein module for its cognate peptide ligand. High-throughput (HTP) analysis of peptide-binding domains by peptide or protein arrays, phage display, mass spectrometry, or other HTP techniques provides new insight into the potential proteinprotein interactions prescribed by individual or even whole families of modules. Systems level analyses have also promoted a deeper understanding of the underlying principles that govern selective proteinprotein interactions and how selectivity evolves. Lastly, there is a growing appreciation for the limitations and potential pitfalls associated with HTP analysis of proteinpeptide interactomes. This review will examine some of the common approaches utilized for large-scale studies of PIDs and suggest a set of standards for the analysis and validation of datasets from large-scale studies of peptide-binding modules. We will also highlight how data from large-scale studies of modular interaction domain families can provide insight into systems level properties such as the linguistics of selective interactions.
RI Liu, Bernard/A-3687-2012;
OI Liu, Bernard/0000-0003-3060-3120; Engelmann, Brett/0000-0002-9845-6668
SN 1615-9853
PD MAY
PY 2012

VL 12
IS 10
BP 1527
EP 1546
DI 10.1002/pmic.201100599
UT WOS:000305474400005
PM 22610655
ER

PT J
AU Kim, T
Tyndel, MS
Huang, HM
Sidhu, SS
Bader, GD
Gfeller, D
Kim, PM
AF Kim, TaeHyung
Tyndel, Marc S.
Huang, Haiming
Sidhu, Sachdev S.
Bader, Gary D.
Gfeller, David
Kim, Philip M.
TI MUSI: an integrated system for identifying multiple specificity from very large peptide or nucleic acid data sets
SO NUCLEIC ACIDS RESEARCH
AB Peptide recognition domains and transcription factors play crucial roles in cellular signaling. They bind linear stretches of amino acids or nucleotides, respectively, with high specificity. Experimental techniques that assess the binding specificity of these domains, such as microarrays or phage display, can retrieve thousands of distinct ligands, providing detailed insight into binding specificity. In particular, the advent of next-generation sequencing has recently increased the throughput of such methods by several orders of magnitude. These advances have helped reveal the presence of distinct binding specificity classes that co-exist within a set of ligands interacting with the same target. Here, we introduce a software system called MUSI that can rapidly analyze very large data sets of binding sequences to determine the relevant binding specificity patterns. Our pipeline provides two major advances. First, it can detect previously unrecognized multiple specificity patterns in any data set. Second, it offers integrated processing of very large data sets from next-generation sequencing machines. The results are visualized as multiple sequence logos describing the different binding preferences of the protein under investigation. We demonstrate the performance of MUSI by analyzing recent phage display data for human SH3 domains as well as microarray data for mouse transcription factors.
SN 0305-1048
PD MAR
PY 2012
VL 40
IS 6
AR e47
DI 10.1093/nar/gkr1294
UT WOS:000302312400008
PM 22210894
ER

PT J
AU Baker, M
AF Baker, Monya
TI THE CHANGES THAT COUNT
SO NATURE
SN 0028-0836
PD FEB 9
PY 2012
VL 482
IS 7384
BP 257
EP 262
UT WOS:000299994100046
PM 22318607
ER

PT J
AU Adkar, BV
Tripathi, A

Sahoo, A
Bajaj, K
Goswami, D
Chakrabarti, P
Swarnkar, MK
Gokhale, RS
Varadarajan, R
AF Adkar, Bharat V.
Tripathi, Arti
Sahoo, Anusmita
Bajaj, Kanika
Goswami, Devrishi
Chakrabarti, Purbani
Swarnkar, Mohit K.
Gokhale, Rajesh S.
Varadarajan, Raghavan
TI Protein Model Discrimination Using Mutational Sensitivity Derived from Deep Sequencing
SO STRUCTURE
AB A major bottleneck in protein structure prediction is the selection of correct models from a pool of decoys. Relative activities of similar to 1,200 individual single-site mutants in a saturation library of the bacterial toxin CcdB were estimated by determining their relative populations using deep sequencing. This phenotypic information was used to define an empirical score for each residue (Rank Score), which correlated with the residue depth, and identify active-site residues. Using these correlations, similar to 98% of correct models of CcdB (RMSD <= 4 angstrom) were identified from a large set of decoys. The model-discrimination methodology was further validated on eleven different monomeric proteins using simulated RankScore values. The methodology is also a rapid, accurate way to obtain relative activities of each mutant in a large pool and derive sequence-structure-function relationships without protein isolation or characterization. It can be applied to any system in which mutational effects can be monitored by a phenotypic readout.
RI Ganju, Shahji/F-3409-2012;
OI Bajaj Pahuja, Kanika/0000-0001-8977-9812
SN 0969-2126
EI 1878-4186
PD FEB 8
PY 2012
VL 20
IS 2
BP 371
EP 381
DI 10.1016/j.str.2011.11.021
UT WOS:000300388000020
PM 22325784
ER

PT J
AU Fowler, DM
Araya, CL
Gerard, W
Fields, S
AF Fowler, Douglas M.
Araya, Carlos L.
Gerard, Wayne
Fields, Stanley
TI Enrich: software for analysis of protein function by enrichment and depletion of variants
SO BIOINFORMATICS
AB Measuring the consequences of mutation in proteins is critical to understanding their function. These measurements are essential in such applications as protein engineering, drug development, protein design and genome sequence analysis. Recently, high-throughput sequencing has been coupled to assays of protein activity, enabling the analysis of large numbers of mutations in parallel. We present Enrich, a tool for analyzing such deep mutational scanning data. Enrich identifies all unique variants (mutants) of a protein in high-throughput sequencing datasets and can correct for sequencing errors using overlapping paired-end reads. Enrich uses the frequency of each variant before and after selection to calculate an enrichment ratio, which is used to estimate fitness. Enrich provides an interactive interface to guide users. It generates user-accessible output for downstream analyses as well as several visualizations of the effects of mutation on function, thereby allowing the user to rapidly quantify and comprehend sequence-function relationships.
OI Araya, Carlos/0000-0002-5512-3062
SN 1367-4803
PD DEC 15
PY 2011
VL 27
IS 24

BP 3430
EP 3431
DI 10.1093/bioinformatics/btr577
UT WOS:000297860000017
PM 22006916
ER

PT J
AU Araya, CL
  Fowler, DM
AF Araya, Carlos L.
  Fowler, Douglas M.
TI Deep mutational scanning: assessing protein function on a massive scale
SO TRENDS IN BIOTECHNOLOGY
AB Analysis of protein mutants is an effective means to understand their function. Protein display is an approach that allows large numbers of mutants of a protein to be selected based on their activity, but only a handful with maximal activity have been traditionally identified for subsequent functional analysis. However, the recent application of high-throughput sequencing (HTS) to protein display and selection has enabled simultaneous assessment of the function of hundreds of thousands of mutants that span the activity range from high to low. Such deep mutational scanning approaches are rapid and inexpensive with the potential for broad utility. In this review, we discuss the emergence of deep mutational scanning, the challenges associated with its use and some of its exciting applications.
OI Araya, Carlos/0000-0002-5512-3062
SN 0167-7799
PD SEP
PY 2011
VL 29
IS 9
BP 435
EP 442
DI 10.1016/j.tibtech.2011.04.003
UT WOS:000294943400003
PM 21561674
ER

PT J
AU Cooper, GM
  Shendure, J
AF Cooper, Gregory M.
  Shendure, Jay
TI Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data
SO NATURE REVIEWS GENETICS
AB Genome and exome sequencing yield extensive catalogues of human genetic variation. However, pinpointing the few phenotypically causal variants among the many variants present in human genomes remains a major challenge, particularly for rare and complex traits wherein genetic information alone is often insufficient. Here, we review approaches to estimate the deleteriousness of single nucleotide variants (SNVs), which can be used to prioritize disease-causal variants. We describe recent advances in comparative and functional genomics that enable systematic annotation of both coding and non-coding variants. Application and optimization of these methods will be essential to find the genetic answers that sequencing promises to hide in plain sight.
RI Sincan, Murat /A-3794-2010;
OI Shendure, Jay/0000-0002-1516-1865
SN 1471-0056
EI 1471-0064
PD SEP
PY 2011
VL 12
IS 9
BP 628
EP 640
DI 10.1038/nrg3046
UT WOS:000294004100009
PM 21850043
ER

PT J
AU Zhang, HK
  Torkamani, A
  Jones, TM
  Ruiz, DI

Pons, J
Lerner, RA
AF Zhang, Hongkai
Torkamani, Ali
Jones, Teresa M.
Ruiz, Diana I.
Pons, Jaume
Lerner, Richard A.
TI Phenotype-information-phenotype cycle for deconvolution of combinatorial antibody libraries selected against complex systems
SO PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA
AB Use of large combinatorial antibody libraries and next-generation sequencing of nucleic acids are two of the most powerful methods in modern molecular biology. The libraries are screened using the principles of evolutionary selection, albeit in real time, to enrich for members with a particular phenotype. This selective process necessarily results in the loss of information about less-fit molecules. On the other hand, sequencing of the library, by itself, gives information that is mostly unrelated to phenotype. If the two methods could be combined, the full potential of very large molecular libraries could be realized. Here we report the implementation of a phenotype-information-phenotype cycle that integrates information and gene recovery. After selection for phage-encoded antibodies that bind to targets expressed on the surface of Escherichia coli, the information content of the selected pool is obtained by pyrosequencing. Sequences that encode specific antibodies are identified by a bioinformatic analysis and recovered by a stringent affinity method that is uniquely suited for gene isolation from a highly degenerate collection of nucleic acids. This approach can be generalized for selection of antibodies against targets that are present as minor components of complex systems.
SN 0027-8424
PD AUG 16
PY 2011
VL 108
IS 33
BP 13456
EP 13461
DI 10.1073/pnas.1111218108
UT WOS:000293895100027
PM 21825149
ER

PT J
AU Smith, CA
Kortemme, T
AF Smith, Colin A.
Kortemme, Tanja
TI Predicting the Tolerated Sequences for Proteins and Protein Interfaces Using RosettaBackrub Flexible Backbone Design
SO PLOS ONE
AB Predicting the set of sequences that are tolerated by a protein or protein interface, while maintaining a desired function, is useful for characterizing protein interaction specificity and for computationally designing sequence libraries to engineer proteins with new functions. Here we provide a general method, a detailed set of protocols, and several benchmarks and analyses for estimating tolerated sequences using flexible backbone protein design implemented in the Rosetta molecular modeling software suite. The input to the method is at least one experimentally determined three-dimensional protein structure or high-quality model. The starting structure(s) are expanded or refined into a conformational ensemble using Monte Carlo simulations consisting of backrub backbone and side chain moves in Rosetta. The method then uses a combination of simulated annealing and genetic algorithm optimization methods to enrich for low-energy sequences for the individual members of the ensemble. To emphasize certain functional requirements (e.g. forming a binding interface), interactions between and within parts of the structure (e.g. domains) can be reweighted in the scoring function. Results from each backbone structure are merged together to create a single estimate for the tolerated sequence space. We provide an extensive description of the protocol and its parameters, all source code, example analysis scripts and three tests applying this method to finding sequences predicted to stabilize proteins or protein interfaces. The generality of this method makes many other applications possible, for example stabilizing interactions with small molecules, DNA, or RNA. Through the use of within-domain reweighting and/or multistate design, it may also be possible to use this method to find sequences that stabilize particular protein conformations or binding interactions over others.
RI Smith, Colin/E-5713-2012
OI Smith, Colin/0000-0002-4651-167X
SN 1932-6203
PD JUL 18
PY 2011
VL 6
IS 7
AR e20451
DI 10.1371/journal.pone.0020451
UT WOS:000292812400003
PM 21789164
ER

PT J
AU Raveh, B
London, N
Zimmerman, L
Schueler-Furman, O
AF Raveh, Barak
London, Nir
Zimmerman, Lior
Schueler-Furman, Ora
TI Rosetta FlexPepDockab-initio: Simultaneous Folding, Docking and Refinement of Peptides onto Their Receptors
SO PLOS ONE
AB Flexible peptides that fold upon binding to another protein molecule mediate a large number of regulatory interactions in the living cell and may provide highly specific recognition modules. We present Rosetta FlexPepDockab-initio, a protocol for simultaneous docking and de-novo folding of peptides, starting from an approximate specification of the peptide binding site. Using the Rosetta fragments library and a coarse-grained structural representation of the peptide and the receptor, FlexPepDockab-initio samples efficiently and simultaneously sampled the space of possible peptide backbone conformations and rigid-body orientations over the receptor surface of a given binding site. The subsequent all-atom refinement of the coarse-grained models includes full side-chain modeling of both the receptor and the peptide, resulting in high-resolution models in which key side-chain interactions are recapitulated. The protocol was applied to a benchmark in which peptides were modeled over receptors in either their bound backbone conformations or in their free, unbound form. Near-native peptide conformations were identified in 18/26 of the bound cases and 7/14 of the unbound cases. The protocol performswell on peptides from various classes of secondary structures, including coiled peptides with unusual turns and kinks. The results presented here significantly extend the scope of state-of-the-art methods for high-resolution peptide modeling, which can now be applied to a wide variety of peptide-protein interactions where no prior information about the peptide backbone conformation is available, enabling detailed structure-based studies and manipulation of those interactions.
SN 1932-6203
PD APR 29
PY 2011
VL 6
IS 4
AR e18934
DI 10.1371/journal.pone.0018934
UT WOS:000290024700038
PM 21572516
ER

PT J
AU Gfeller, D
Butty, F
Wierzbicka, M
Verschueren, E
Vanhee, P
Huang, HM
Ernst, A
Dar, N
Stagljar, I
Serrano, L
Sidhu, SS
Bader, GD
Kim, PM
AF Gfeller, David
Butty, Frank
Wierzbicka, Marta
Verschueren, Erik
Vanhee, Peter
Huang, Haiming
Ernst, Andreas
Dar, Nisa
Stagljar, Igor
Serrano, Luis
Sidhu, Sachdev S.
Bader, Gary D.
Kim, Philip M.
TI The multiple-specificity landscape of modular peptide recognition domains
SO MOLECULAR SYSTEMS BIOLOGY
AB Modular protein interaction domains form the building blocks of eukaryotic signaling pathways. Many of them, known as peptide recognition domains, mediate protein interactions by recognizing short, linear amino acid stretches on the surface of their cognate partners with high specificity.

Residues in these stretches are usually assumed to contribute independently to binding, which has led to a simplified understanding of protein interactions. Conversely, we observe in large binding peptide data sets that different residue positions display highly significant correlations for many domains in three distinct families (PDZ, SH3 and WW). These correlation patterns reveal a widespread occurrence of multiple binding specificities and give novel structural insights into protein interactions. For example, we predict a new binding mode of PDZ domains and structurally rationalize it for DLG1 PDZ1. We show that multiple specificity more accurately predicts protein interactions and experimentally validate some of the predictions for the human proteins DLG1 and SCRIB. Overall, our results reveal a rich specificity landscape in peptide recognition domains, suggesting new ways of encoding specificity in protein interaction networks. Molecular Systems Biology 7: 484; published online 26 April 2011; doi:10.1038/msb.2011.18
RI Bader, Gary/C-1176-2009; Serrano, Luis/B-3355-2013; verschueren, erik/D-4436-2015
OI Bader, Gary/0000-0003-0185-8861; Serrano, Luis/0000-0002-5276-1392; verschueren, erik/0000-0001-5842-6344
SN 1744-4292
PD APR
PY 2011
VL 7
AR 484
DI 10.1038/msb.2011.18
UT WOS:000290411600005
PM 21525870
ER

PT J
AU Caberoy, NB
Alvarado, G
Li, W
AF Caberoy, Nora B.
Alvarado, Gabriela
Li, Wei
TI Identification of Calpain Substrates by ORF Phage Display
SO MOLECULES
AB Substrate identification is the key to defining molecular pathways or cellular processes regulated by proteases. Although phage display with random peptide libraries has been used to analyze substrate specificity of proteases, it is difficult to deduce endogenous substrates from mapped peptide motifs. Phage display with conventional cDNA libraries identifies high percentage of non-open reading frame (non-ORF) clones, which encode short unnatural peptides, owing to uncontrollable reading frames of cellular proteins. We recently developed ORF phage display to identify endogenous proteins with specific binding or functional activity with minimal reading frame problem. Here we used calpain 2 as a protease to demonstrate that ORF phage display is capable of identifying endogenous substrates and showed its advantage to re-verify and characterize the identified substrates without requiring pure substrate proteins. An ORF phage display cDNA library with C-terminal biotin was bound to immobilized streptavidin and released by cleavage with calpain 2. After three rounds of phage selection, eleven substrates were identified, including calpastatin of endogenous calpain inhibitor. These results suggest that ORF phage display is a valuable technology to identify endogenous substrates for proteases.
SN 1420-3049
PD FEB
PY 2011
VL 16
IS 2
BP 1739
EP 1748
DI 10.3390/molecules16021739
UT WOS:000287745400048
PM 21339709
ER

PT J
AU Baker, M
AF Baker, Monya
TI Stan Fields and Doug Fowler
SO NATURE METHODS
SN 1548-7091
PD SEP
PY 2010
VL 7
IS 9
BP 663
EP 663
DI 10.1038/nmeth0910-663
UT WOS:000281429200002
ER

EF
`

Green

Thursday, September 29, 2016 2:47 PM

TI Causes of evolutionary rate variation among protein sites
SO NATURE REVIEWS GENETICS
AB It has long been recognized that certain sites within a protein, such as sites in the protein core or catalytic residues in enzymes, are evolutionarily more conserved than other sites. However, our understanding of rate variation among sites remains surprisingly limited. Recent progress to address this includes the development of a wide array of reliable methods to estimate site-specific substitution rates from sequence alignments. In addition, several molecular traits have been identified that correlate with site-specific mutation rates, and novel mechanistic biophysical models have been proposed to explain the observed correlations. Nonetheless, current models explain, at best, approximately 60% of the observed variance, highlighting the limitations of current methods and models and the need for new research directions.

TI Massively parallel enzyme kinetics reveals the substrate recognition landscape of the metalloprotease ADAMTS13
SO PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA
AB Proteases play important roles in many biologic processes and are key mediators of cancer, inflammation, and thrombosis. However, comprehensive and quantitative techniques to define the substrate specificity profile of proteases are lacking. The metalloprotease ADAMTS13 regulates blood coagulation by cleaving vonWillebrand factor (VWF), reducing its procoagulant activity. A mutagenized substrate phage display library based on a 73-amino acid fragment of VWF was constructed, and the ADAMTS13-dependent change in library complexity was evaluated over reaction time points, using high-throughput sequencing. Reaction rate constants (k(cat)/K-M) were calculated for nearly every possible single amino acid substitution within this fragment. This massively parallel enzyme kinetics analysis detailed the specificity of ADAMTS13 and demonstrated the critical importance of the P1-P1' substrate residues while defining exosite binding domains. These data provided empirical evidence for the propensity for epistasis within VWF and showed strong correlation to conservation across orthologs, highlighting evolutionary selective pressures for VWF.

TI Massively Parallel Functional Analysis of BRCA1 RING Domain Variants
SO GENETICS
AB Interpreting variants of uncertain significance (VUS) is a central challenge in medical genetics. One approach is to experimentally measure the functional consequences of VUS, but to date this approach has been post hoc and low throughput. Here we use massively parallel assays to measure the effects of nearly 2000 missense substitutions in the RING domain of BRCA1 on its E3 ubiquitin ligase activity and its binding to the BARD1 RING domain. From the resulting scores, we generate a model to predict the capacities of full-length BRCA1 variants to support homology-directed DNA repair, the essential role of BRCA1 in tumor suppression, and show that it outperforms widely used biological-effect prediction algorithms. We envision that massively parallel functional assays may facilitate the prospective interpretation of variants observed in clinical sequencing.

TI High-Resolution Sequence-Function Mapping of Full-Length Proteins
SO PLOS ONE
AB Comprehensive sequence-function mapping involves detailing the fitness contribution of every possible single mutation to a gene by comparing the abundance of each library variant before and after

selection for the phenotype of interest. Deep sequencing of library DNA allows frequency reconstruction for tens of thousands of variants in a single experiment, yet short read lengths of current sequencers makes it challenging to probe genes encoding full-length proteins. Here we extend the scope of sequence-function maps to entire protein sequences with a modular, universal sequence tiling method. We demonstrate the approach with both growth-based selections and FACS screening, offer parameters and best practices that simplify design of experiments, and present analytical solutions to normalize data across independent selections. Using this protocol, sequence-function maps covering full sequences can be obtained in four to six weeks. Best practices introduced in this manuscript are fully compatible with, and complementary to, other recently published sequence-function mapping protocols.

TI Combining Natural Sequence Variation with High Throughput Mutational Data to Reveal Protein Interaction Sites
SO PLOS GENETICS
AB Many protein interactions are conserved among organisms despite changes in the amino acid sequences that comprise their contact sites, a property that has been used to infer the location of these sites from protein homology. In an inter-species complementation experiment, a sequence present in a homologue is substituted into a protein and tested for its ability to support function. Therefore, substitutions that inhibit function can identify interaction sites that changed over evolution. However, most of the sequence differences within a protein family remain unexplored because of the small-scale nature of these complementation approaches. Here we use existing high throughput mutational data on the in vivo function of the RRM2 domain of the Saccharomyces cerevisiae poly(A)-binding protein, Pab1, to analyze its sites of interaction. Of 197 single amino acid differences in 52 Pab1 homologues, 17 reduce the function of Pab1 when substituted into the yeast protein. The majority of these deleterious mutations interfere with the binding of the RRM2 domain to eIF4G1 and eIF4G2, isoforms of a translation initiation factor. A large-scale mutational analysis of the RRM2 domain in a two-hybrid assay for eIF4G1 binding supports these findings and identifies peripheral residues that make a smaller contribution to eIF4G1 binding. Three single amino acid substitutions in yeast Pab1 corresponding to residues from the human orthologue are deleterious and eliminate binding to the yeast eIF4G isoforms. We create a triple mutant that carries these substitutions and other humanizing substitutions that collectively support a switch in binding specificity of RRM2 from the yeast eIF4G1 to its human orthologue. Finally, we map other deleterious substitutions in Pab1 to inter-domain (RRM2-RRM1) or protein-RNA (RRM2-poly(A)) interaction sites. Thus, the combined approach of large-scale mutational data and evolutionary conservation can be used to characterize interaction sites at single amino acid resolution.

Fitness analyses of all possible point mutations for regions of genes in yeast

Relevant
Maybe
Not Sure
Not Relevant

Wednesday, September 28, 2016 5:41 PM

Cited by 20:

FN Thomson Reuters Web of Science™
VR 1.0
PT J
AU Abriata, LA
Bovigny, C
Dal Peraro, M
AF Abriata, Luciano A.
Bovigny, Christophe
Dal Peraro, Matteo
TI Detection and sequence/structure mapping of biophysical constraints to protein variation in saturated mutational libraries and protein sequence alignments with a dedicated server
SO BMC BIOINFORMATICS
AB Background: Protein variability can now be studied by measuring high-resolution tolerance-to-substitution maps and fitness landscapes in saturated mutational libraries. But these rich and expensive datasets are typically interpreted coarsely, restricting detailed analyses to positions of extremely high or low variability or dubbed important beforehand based on existing knowledge about active sites, interaction surfaces, (de) stabilizing mutations, etc.
Results: Our new webserver PsychoProt (freely available without registration at http://psychoprot.epfl.ch or at http://lucianoabriata.altervista.org/psychoprot/index.html) helps to detect, quantify, and sequence/structure map the biophysical and biochemical traits that shape amino acid preferences throughout a protein as determined by deep-sequencing of saturated mutational libraries or from large alignments of naturally occurring variants.
Discussion: We exemplify how PsychoProt helps to (i) unveil protein structure-function relationships from experiments and from alignments that are consistent with structures according to coevolution analysis, (ii) recall global information about structural and functional features and identify hitherto unknown constraints to variation in alignments, and (iii) point at different sources of variation among related experimental datasets or between experimental and alignment-based data. Remarkably, metabolic costs of the amino acids pose strong constraints to variability at protein surfaces in nature but not in the laboratory. This and other differences call for caution when extrapolating results from in vitro experiments to natural scenarios in, for example, studies of protein evolution.
Conclusion: We show through examples how PsychoProt can be a useful tool for the broad communities of structural biology and molecular evolution, particularly for studies about protein modeling, evolution and design.
SN 1471-2105
PD JUN 17
PY 2016
VL 17
AR 242
DI 10.1186/s12859-016-1124-4
UT WOS:000378845300001
PM 27315797
ER

PT J
AU Mayor, D
Barlow, K
Thompson, S
Barad, BA
Bonny, AR
Cario, CL
Gaskins, G
Liu, ZR
Deming, L

Axen, SD
Caceres, E
Chen, W
Cuesta, A
Gate, RE
Green, EM
Hulce, KR
Ji, WY
Kenner, LR
Mensa, B
Morinishi, LS
Moss, SM
Mravic, M
Muir, RK
Niekamp, S
Nnadi, CI
Palovcak, E
Poss, EM
Ross, TD
Salcedo, EC
See, SK
Subramaniam, M
Wong, AW
Li, J
Thorn, KS
Conchuir, SO
Roscoe, BP
Chow, ED
DeRisi, JL
Kortemme, T
Bolon, DN
Fraser, JS
AF Mayor, David
Barlow, Kyle
Thompson, Samuel
Barad, Benjamin A.
Bonny, Alain R.
Cario, Clinton L.
Gaskins, Garrett
Liu, Zairan
Deming, Laura
Axen, Seth D.
Caceres, Elena
Chen, Weilin
Cuesta, Adolfo
Gate, Rachel E.
Green, Evan M.
Hulce, Kaitlin R.
Ji, Weiyue
Kenner, Lillian R.
Mensa, Bruk
Morinishi, Leanna S.
Moss, Steven M.
Mravic, Marco
Muir, Ryan K.
Niekamp, Stefan
Nnadi, Chimno I.
Palovcak, Eugene

Poss, Erin M.
Ross, Tyler D.
Salcedo, Eugenia C.
See, Stephanie K.
Subramaniam, Meena
Wong, Allison W.
Li, Jennifer
Thorn, Kurt S.
Conchuir, Shane O.
Roscoe, Benjamin P.
Chow, Eric D.
DeRisi, Joseph L.
Kortemme, Tanja
Bolon, Daniel N.
Fraser, James S.
TI Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting
SO ELIFE
AB Ubiquitin is essential for eukaryotic life and varies in only 3 amino acid positions between yeast and humans. However, recent deep sequencing studies indicate that ubiquitin is highly tolerant to single mutations. We hypothesized that this tolerance would be reduced by chemically induced physiologic perturbations. To test this hypothesis, a class of first year UCSF graduate students employed deep mutational scanning to determine the fitness landscape of all possible single residue mutations in the presence of five different small molecule perturbations. These perturbations uncover 'shared sensitized positions' localized to areas around the hydrophobic patch and the C-terminus. In addition, we identified perturbation specific effects such as a sensitization of His68 in HU and a tolerance to mutation at Lys63 in DTT. Our data show how chemical stresses can reduce buffering effects in the ubiquitin proteasome system. Finally, this study demonstrates the potential of lab-based interdisciplinary graduate curriculum.

eLife digest The ability of an organism to grow and reproduce, that is, it's "fitness", is determined by how its genes interact with the environment. Yeast is a model organism in which researchers can control the exact mutations present in the yeast's genes (its genotype) and the conditions in which the yeast cells live (their environment). This allows researchers to measure how a yeast cell's genotype and environment affect its fitness.

Ubiquitin is a protein that many organisms depend on to manage cell stress by acting as a tag that targets other proteins for degradation. Essential proteins such as ubiquitin often remain unchanged by mutation over long periods of time. As a result, these proteins evolve very slowly. Like all proteins, ubiquitin is built from a chain of amino acid molecules linked together, and the ubiquitin proteins of yeast and humans are made of almost identical sequences of amino acids.

Although ubiquitin has barely changed its sequence over evolution, previous studies have shown that under normal growth conditions in the laboratory most amino acids in ubiquitin can be mutated without any loss of cell fitness. This led Mayor et al. to hypothesize that treating the yeast cells with chemicals that cause cell stress might lead to amino acids in ubiquitin becoming more sensitive to mutation.

To test this idea, a class of graduate students at the University of California, San Francisco grew yeast cells with different ubiquitin mutations together, and with different chemicals that induce cell stress, and measured their growth rates. Sequencing the ubiquitin gene in the thousands of tested yeast cells revealed that three of the chemicals cause a shared set of amino acids in ubiquitin to become more sensitive to mutation.

This result suggests that these amino acids are important for the stress response, possibly by altering the ability of yeast cells to target certain proteins for degradation. Conversely, another chemical causes yeast to become more tolerant to changes in the ubiquitin sequence. The experiments also link changes in particular amino acids in ubiquitin to specific stress responses.

Mayor et al. show that many of ubquitin's amino acids are sensitive to mutation under different stress conditions, while others can be mutated to form different amino acids without effecting fitness. By testing the effects of other chemicals, future experiments could further characterize how the yeast's genotype and environment interact.
OI Mavor, David/0000-0002-8281-1493; Barlow, Kyle/0000-0002-9787-0066
SN 2050-084X
PD APR 25
PY 2016
VL 5
AR e15802
DI 10.7554/eLife.15802

UT WOS:000376605700001
ER

PT J
AU Mishra, P
Flynn, JM
Starr, TN
Bolon, DNA
AF Mishra, Parul
Flynn, Julia M.
Starr, Tyler N.
Bolon, Daniel N. A.
TI Systematic Mutant Analyses Elucidate General and Client-Specific Aspects
of Hsp90 Function
SO CELL REPORTS
AB To probe the mechanism of the Hsp90 chaperone that is required for the maturation of many signaling proteins in eukaryotes, we analyzed the effects of all individual amino acid changes in the ATPase domain on yeast growth rate. The sensitivity of a position to mutation was strongly influenced by proximity to the phosphates of ATP, indicating that ATPase-driven conformational changes impose stringent physical constraints on Hsp90. To investigate how these constraints may vary for different clients, we performed biochemical analyses on a panel of Hsp90 mutants spanning the full range of observed fitness effects. We observed distinct effects of nine Hsp90 mutations on activation of v-src and glucocorticoid receptor (GR), indicating that different chaperone mechanisms can be utilized for these clients. These results provide a detailed guide for understanding Hsp90 mechanism and highlight the potential for inhibitors of Hsp90 that target a subset of clients.
SN 2211-1247
PD APR 19
PY 2016
VL 15
IS 3
BP 588
EP 598
DI 10.1016/j.celrep.2016.03.046
UT WOS:000374498900014
PM 27068472
ER

PT J
AU Phillips, AM
Shoulders, MD
AF Phillips, Angela M.
Shoulders, Matthew D.
TI The Path of Least Resistance: Mechanisms to Reduce Influenza's
Sensitivity to Oseltamivir
SO JOURNAL OF MOLECULAR BIOLOGY
SN 0022-2836
EI 1089-8638
PD FEB 13
PY 2016
VL 428
IS 3
BP 533
EP 537
DI 10.1016/j.jmb.2015.12.019
UT WOS:000371839800001
PM 26748011
ER

PT J

AU Jiang, L
Liu, P
Bank, C
Renzette, N
Prachanronarong, K
Yilmaz, LS
Caffrey, DR
Zeldovich, KB
Schiffer, CA
Kowalik, TF
Jensen, JD
Finberg, RW
Wang, JP
Bolon, DNA
AF Jiang, Li
Liu, Ping
Bank, Claudia
Renzette, Nicholas
Prachanronarong, Kristina
Yilmaz, Lutfu S.
Caffrey, Daniel R.
Zeldovich, Konstantin B.
Schiffer, Celia A.
Kowalik, Timothy F.
Jensen, Jeffrey D.
Finberg, Robert W.
Wang, Jennifer P.
Bolon, Daniel N. A.
TI A Balance between Inhibitor Binding and Substrate Processing Confers
Influenza Drug Resistance
SO JOURNAL OF MOLECULAR BIOLOGY
AB The therapeutic benefits of the neuraminidase (NA) inhibitor oseltamivir are dampened by the emergence of drug resistance mutations in influenza A virus (IAV). To investigate the mechanistic features that underlie resistance, we developed an approach to quantify the effects of all possible single-nucleotide substitutions introduced into important regions of NA. We determined the experimental fitness effects of 450 nucleotide mutations encoding positions both surrounding the active site and at more distant sites in an N1 strain of IAV in the presence and absence of oseltamivir. NA mutations previously known to confer oseltamivir resistance in N1 strains, including H275Y and N295S, were adaptive in the presence of drug, indicating that our experimental system captured salient features of real-world selection pressures acting on NA. We identified mutations, including several at position 223, that reduce the apparent affinity for oseltamivir in vitro. Position 223 of NA is located adjacent to a hydrophobic portion of oseltamivir that is chemically distinct from the substrate, making it a hotspot for substitutions that preferentially impact drug binding relative to substrate processing. Furthermore, two NA mutations, K221N and Y276F, each reduce susceptibility to oseltamivir by increasing NA activity without altering drug binding. These results indicate that competitive expansion of IAV in the face of drug pressure is mediated by a balance between inhibitor binding and substrate processing.
SN 0022-2836
EI 1089-8638
PD FEB 13
PY 2016
VL 428
IS 3
BP 538
EP 553
DI 10.1016/j.jmb.2015.11.027
UT WOS:000371839800002
PM 26656922
ER

PT J
AU Foight, GW
Keating, AE
AF Foight, Glenna Wink
Keating, Amy E.
TI Locating Herpesvirus Bcl-2 Homologs in the Specificity Landscape of
Anti-Apoptotic Bcl-2 Proteins
SO JOURNAL OF MOLECULAR BIOLOGY
AB Viral homologs of the anti-apoptotic Bcl-2 proteins are highly diverged from their mammalian counterparts, yet they perform overlapping functions by binding and inhibiting BH3 (Bcl-2 homology 3)-motif-containing proteins. We investigated the BH3 binding properties of the herpesvirus Bcl-2 homologs KSBcl-2, BHRF1, and M11, as they relate to those of the human Bcl-2 homologs Mcl-1, Bfl-1, Bcl-w, Bcl-x(L), and Bcl-2. Analysis of the sequence and structure of the BH3 binding grooves showed that, despite low sequence identity, M11 has structural similarities to Bcl-x(L), Bcl-2, and Bcl-w. BHRF1 and KSBcl-2 are more structurally similar to Mcl-1 than to the other human proteins. Binding to human BH3-like peptides showed that KSBcl-2 has similar specificity to Mcl-1, and BHRF1 has a restricted binding profile; M11 binding preferences are distinct from those of Bcl-x(L), Bcl-2, and Bcl-w. Because KSBcl-2 and BHRF1 are from human herpesviruses associated with malignancies, we screened computationally designed BH3 peptide libraries using bacterial surface display to identify selective binders of KSBcl-2 or BHRF1. The resulting peptides bound to KSBcl-2 and BHRF1 in preference to Bfl-1, Bcl-w, Bcl-x(L), and Bcl-2 but showed only modest specificity over Mcl-1. Rational mutagenesis increased specificity against Mcl-1, resulting in a peptide with a dissociation constant of 2.9 nM for binding to KSBcl-2 and >1000-fold specificity over other Bcl-2 proteins, as well as a peptide with >70-fold specificity for BHRF1. In addition to providing new insights into viral Bcl-2 binding specificity, this study will inform future work analyzing the interaction properties of homologous binding domains and designing specific protein interaction partners. (C) 2015 Elsevier Ltd. All rights reserved.
RI Foight, Glenna/J-6032-2015
OI Foight, Glenna/0000-0003-3749-7092
SN 0022-2836
EI 1089-8638
PD JUL 31
PY 2015
VL 427
IS 15
BP 2468
EP 2490
DI 10.1016/j.jmb.2015.05.015
UT WOS:000359169400006
PM 26009469
ER

PT J
AU Reich, L
Dutta, S
Keating, AE
AF Reich, Lothar Luther
Dutta, Sanjib
Keating, Amy E.
TI SORTCERY-A High-Throughput Method to Affinity Rank Peptide Ligands
SO JOURNAL OF MOLECULAR BIOLOGY
AB Uncovering the relationships between peptide and protein sequences and binding properties is critical for successfully predicting, re-designing and inhibiting protein-protein interactions. Systematically collected data that link protein sequence to binding are valuable for elucidating determinants of protein interaction but are rare in the literature because such data are experimentally difficult to generate. Here we describe SORTCERY, a high-throughput method that we have used to rank hundreds of yeast-displayed peptides according to their affinities for a target interaction partner. The procedure involves fluorescence-activated cell sorting of a library, deep sequencing of sorted pools and downstream computational analysis. We have developed theoretical models and statistical tools that assist in planning these stages. We demonstrate SORTCERY's utility by ranking 1026 BH3 (Bcl-2 homology 3) peptides with respect to their affinities for the anti-apoptotic protein Bcl-x(L). Our results are in striking agreement with measured affinities for 19 individual peptides with dissociation constants ranging from 0.1 to 60 nM. High-resolution ranking can be used to improve our

understanding of sequence-function relationships and to support the development of computational models for predicting and designing novel interactions. (C) 2014 Elsevier Ltd. All rights reserved.
SN 0022-2836
EI 1089-8638
PD JUN 5
PY 2015
VL 427
IS 11
SI SI
BP 2135
EP 2150
DI 10.1016/j.jmb.2014.09.025
UT WOS:000355028100010
PM 25311858
ER

PT J
AU Van Blarcom, T
Rossi, A
Foletti, D
Sundar, P
Pitts, S
Bee, C
Witt, JM
Melton, Z
Hasa-Moreno, A
Shaughnessy, L
Telman, D
Zhao, L
Cheung, WL
Berka, J
Zhai, WW
Strop, P
Chaparro-Riggers, J
Shelton, DL
Pons, J
Rajpal, A
AF Van Blarcom, Thomas
Rossi, Andrea
Foletti, Davide
Sundar, Purnima
Pitts, Steven
Bee, Christine
Witt, Jody Melton
Melton, Zea
Hasa-Moreno, Adela
Shaughnessy, Lee
Telman, Dilduz
Zhao, Lora
Cheung, Wai Ling
Berka, Jan
Zhai, Wenwu
Strop, Pavel
Chaparro-Riggers, Javier
Shelton, David L.
Pons, Jaume
Rajpal, Arvind
TI Precise and Efficient Antibody Epitope Determination through Library

Design, Yeast Display and Next-Generation Sequencing
SO JOURNAL OF MOLECULAR BIOLOGY
AB The ability of antibodies to bind an antigen with a high degree of affinity and specificity has led them to become the largest and fastest growing class of therapeutic proteins. Clearly identifying the epitope at which they bind their cognate antigen provides insight into their mechanism of action and helps differentiate antibodies that bind the same antigen. Here, we describe a method to precisely and efficiently map the epitopes of a panel of antibodies in parallel over the course of several weeks. This method relies on the combination of rational library design, quantitative yeast surface display and next-generation DNA sequencing and was demonstrated by mapping the epitopes of several antibodies that neutralize alpha toxin from Staphylococcus aureus. The accuracy of this method was confirmed by comparing the results to the co-crystal structure of one antibody and alpha toxin and was further refined by the inclusion of a lower-affinity variant of the antibody. In addition, this method produced quantitative insight into the epitope residues most critical for the antibody antigen interaction and enabled the relative affinities of each antibody toward alpha toxin variants to be estimated. This affinity estimate serves as a predictor of neutralizing antibody potency and was used to anticipate the ability of each antibody to effectively bind and neutralize naturally occurring alpha toxin variants secreted by strains of S. aureus, including clinically relevant strains. Ultimately this type information can be used to help select the best clinical candidate among a set of antibodies against a given antigen. (C) 2014 Elsevier Ltd. All rights reserved.
SN 0022-2836
EI 1089-8638
PD MAR 27
PY 2015
VL 427
IS 6
BP 1513
EP 1534
DI 10.1016/j.jmb.2014.09.020
PN B
UT WOS:000351798700022
PM 25284753
ER

PT J
AU Kowalsky, CA
Klesmith, JR
Stapleton, JA
Kelly, V
Reichkitzer, N
Whitehead, TA
AF Kowalsky, Caitlin A.
Klesmith, Justin R.
Stapleton, James A.
Kelly, Vince
Reichkitzer, Nolan
Whitehead, Timothy A.
TI High-Resolution Sequence-Function Mapping of Full-Length Proteins
SO PLOS ONE
AB Comprehensive sequence-function mapping involves detailing the fitness contribution of every possible single mutation to a gene by comparing the abundance of each library variant before and after selection for the phenotype of interest. Deep sequencing of library DNA allows frequency reconstruction for tens of thousands of variants in a single experiment, yet short read lengths of current sequencers makes it challenging to probe genes encoding full-length proteins. Here we extend the scope of sequence-function maps to entire protein sequences with a modular, universal sequence tiling method. We demonstrate the approach with both growth-based selections and FACS screening, offer parameters and best practices that simplify design of experiments, and present analytical solutions to normalize data across independent selections. Using this protocol, sequence-function maps covering full sequences can be obtained in four to six weeks. Best practices introduced in this manuscript are fully compatible with, and complementary to, other recently published sequence-function mapping protocols.
OI Klesmith, Justin/0000-0003-2908-9355
SN 1932-6203
PD MAR 19

PY 2015
VL 10
IS 3
AR e0118193
DI 10.1371/journal.pone.0118193
UT WOS:000351425400025
PM 25790064
ER

PT J
AU Abriata, LA
  Palzkill, T
  Dal Peraro, M
AF Abriata, Luciano A.
  Palzkill, Timothy
  Dal Peraro, Matteo
TI How Structural and Physicochemical Determinants Shape Sequence
  Constraints in a Functional Enzyme
SO PLOS ONE
AB The need for interfacing structural biology and biophysics to molecular evolution is being increasingly recognized. One part of the big problem is to understand how physics and chemistry shape the sequence space available to functional proteins, while satisfying the needs of biology. Here we present a quantitative, structure-based analysis of a high-resolution map describing the tolerance to all substitutions in all positions of a functional enzyme, namely a TEM lactamase previously studied through deep sequencing of mutants growing in competition experiments with selection against ampicillin. Substitutions are rarely observed within 7 angstrom of the active site, a stringency that is relaxed slowly and extends up to 15-20 20 angstrom, with buried residues being especially sensitive. Substitution patterns in over one third of the residues can be quantitatively modeled by monotonic dependencies on amino acid descriptors and predictions of changes in folding stability. Amino acid volume and steric hindrance shape constraints on the protein core; hydrophobicity and solubility shape constraints on hydrophobic clusters underneath the surface, and on salt bridges and polar networks at the protein surface together with charge and hydrogen bonding capacity. Amino acid solubility, flexibility and conformational descriptors also provide additional constraints at many locations. These findings provide fundamental insights into the chemistry underlying protein evolution and design, by quantitating links between sequence and different protein traits, illuminating subtle and unexpected sequence-trait relationships and pinpointing what traits are sacrificed upon gain-of-function mutation.
SN 1932-6203
PD FEB 23
PY 2015
VL 10
IS 2
AR e0118684
DI 10.1371/journal.pone.0118684
UT WOS:000350662100201
PM 25706742
ER

PT J
AU Bank, C
  Hietpas, RT
  Jensen, JD
  Bolon, DNA
AF Bank, Claudia
  Hietpas, Ryan T.
  Jensen, Jeffrey D.
  Bolon, Daniel N. A.
TI A Systematic Survey of an Intragenic Epistatic Landscape
SO MOLECULAR BIOLOGY AND EVOLUTION
AB Mutations are the source of evolutionary variation. The interactions of multiple mutations can have important effects on fitness and evolutionary trajectories. We have recently described the distribution of fitness effects of all single

mutations for a nine-amino-acid region of yeast Hsp90 (Hsp82) implicated in substrate binding. Here, we report and discuss the distribution of intragenic epistatic effects within this region in seven Hsp90 point mutant backgrounds of neutral to slightly deleterious effect, resulting in an analysis of more than 1,000 double mutants. We find negative epistasis between substitutions to be common, and positive epistasis to be rare-resulting in a pattern that indicates a drastic change in the distribution of fitness effects one step away from the wild type. This can be well explained by a concave relationship between phenotype and genotype (i.e., a concave shape of the local fitness landscape), suggesting mutational robustness intrinsic to the local sequence space. Structural analyses indicate that, in this region, epistatic effects are most pronounced when a solvent-inaccessible position is involved in the interaction. In contrast, all 18 observations of positive epistasis involved at least one mutation at a solvent-exposed position. By combining the analysis of evolutionary and biophysical properties of an epistatic landscape, these results contribute to a more detailed understanding of the complexity of protein evolution.
OI Bank, Claudia/0000-0003-4730-758X; Bolon, Daniel/0000-0001-5857-6676
SN 0737-4038
EI 1537-1719
PD JAN
PY 2015
VL 32
IS 1
BP 229
EP 238
DI 10.1093/molbev/msu301
UT WOS:000350050200021
PM 25371431
ER

PT J
AU Roscoe, BP
  Bolon, DNA
AF Roscoe, Benjamin P.
  Bolon, Daniel N. A.
TI Systematic Exploration of Ubiquitin Sequence, E1 Activation Efficiency,
  and Experimental Fitness in Yeast
SO JOURNAL OF MOLECULAR BIOLOGY
AB The complexity of biological interaction networks poses a challenge to understanding the function of individual connections in the overall network. To address this challenge, we developed a high-throughput reverse engineering strategy to analyze how thousands of specific perturbations (encompassing all point mutations in a central gene) impact both a specific edge (interaction to a directly connected node) and an overall network function. We analyzed the effects of ubiquitin mutations on activation by the E1 enzyme and compared these to effects on yeast growth rate. Using this approach, we delineated ubiquitin mutations that selectively impacted the ubiquitin-E1 edge. We find that the elasticity function relating the efficiency of ubiquitin-E1 interaction to growth rate is non-linear and that a greater than 50-fold decrease in E1 activation efficiency is required to reduce growth rate by 2-fold. Despite the robustness of fitness to decreases in E1 activation efficiency, the effects of most ubiquitin mutations on El activation paralleled the effects on growth rate. Our observations indicate that most ubiquitin mutations that disrupt E1 activation also disrupt other functions. The structurally characterized ubiquitin-E1 interlace encompasses the interlaces of ubiquitin with most other known binding partners, and we propose that this enables E1 in wild-type cells to selectively activate ubiquitin protein molecules capable of binding to other partners from the cytoplasmic pool of ubiquitin protein that will include molecules with chemical damage and/or errors from transcription and translation. (C) 2014 Elsevier Ltd. All rights reserved.
OI Bolon, Daniel/0000-0001-5857-6676
SN 0022-2836
EI 1089-8638
PD JUL 29
PY 2014
VL 426
IS 15
BP 2854
EP 2870
DI 10.1016/j.jmb.2014.05.019

UT WOS:000340302700012
PM 24862281
ER

PT J
AU Bank, C
Hietpas, RT
Wong, A
Bolon, DN
Jensen, JD
AF Bank, Claudia
Hietpas, Ryan T.
Wong, Alex
Bolon, Daniel N.
Jensen, Jeffrey D.
TI A Bayesian MCMC Approach to Assess the Complete Distribution of Fitness
Effects of New Mutations: Uncovering the Potential for Adaptive Walks in
Challenging Environments
SO GENETICS
AB The role of adaptation in the evolutionary process has been contentious for decades. At the heart of the century-old debate between neutralists and selectionists lies the distribution of fitness effects (DFE)that is, the selective effect of all mutations. Attempts to describe the DFE have been varied, occupying theoreticians and experimentalists alike. New high-throughput techniques stand to make important contributions to empirical efforts to characterize the DFE, but the usefulness of such approaches depends on the availability of robust statistical methods for their interpretation. We here present and discuss a Bayesian MCMC approach to estimate fitness from deep sequencing data and use it to assess the DFE for the same 560 point mutations in a coding region of Hsp90 in Saccharomyces cerevisiae across six different environmental conditions. Using these estimates, we compare the differences in the DFEs resulting from mutations covering one-, two-, and three-nucleotide steps from the wild typeshowing that multiple-step mutations harbor more potential for adaptation in challenging environments, but also tend to be more deleterious in the standard environment. All observations are discussed in the light of expectations arising from Fisher's geometric model.
OI Bank, Claudia/0000-0003-4730-758X; Bolon, Daniel/0000-0001-5857-6676
SN 1943-2631
PD MAR
PY 2014
VL 196
IS 3
BP 841
EP +
DI 10.1534/genetics.113.156190
UT WOS:000333905500020
PM 24398421
ER

PT J
AU Tripathi, A
Varadarajan, R
AF Tripathi, Arti
Varadarajan, Raghavan
TI Residue specific contributions to stability and activity inferred from
saturation mutagenesis and deep sequencing
SO CURRENT OPINION IN STRUCTURAL BIOLOGY
AB Multiple methods currently exist for rapid construction and screening of single-site saturation mutagenesis (SSM) libraries in which every codon or nucleotide in a DNA fragment is individually randomized. Nucleotide sequences of each library member before and after screening or selection can be obtained through deep sequencing. The relative enrichment of each mutant at each position provides information on its contribution to protein activity or ligand-binding under the conditions of the screen. Such saturation scans have been applied to diverse proteins to delineate hot-spot residues, stability determinants, and for comprehensive fitness estimates. The data have been used to design proteins with enhanced stability, activity and altered specificity relative to wild-type, to test computational predictions of binding

affinity, and for protein model discrimination. Future improvements in deep sequencing read lengths and accuracy should allow comprehensive studies of epistatic effects, of combinational variation at multiple sites, and identification of spatially proximate residues.
SN 0959-440X
EI 1879-033X
PD FEB
PY 2014
VL 24
BP 63
EP 71
DI 10.1016/j.sbi.2013.12.001
UT WOS:000335100300009
PM 24721454
ER

PT J
AU Wagenaar, TR
Ma, LY
Roscoe, B
Park, SM
Bolon, DN
Green, MR
AF Wagenaar, Timothy R.
Ma, Leyuan
Roscoe, Benjamin
Park, Sung Mi
Bolon, Daniel N.
Green, Michael R.
TI Resistance to vemurafenib resulting from a novel mutation in the
BRAFV600E kinase domain
SO PIGMENT CELL & MELANOMA RESEARCH
AB Resistance to the BRAF inhibitor vemurafenib poses a significant problem for the treatment of BRAFV600E-positive melanomas. It is therefore critical to prospectively identify all vemurafenib resistance mechanisms prior to their emergence in the clinic. The vemurafenib resistance mechanisms described to date do not result from secondary mutations within BRAFV600E. To search for possible mutations within BRAFV600E that can confer drug resistance, we developed a systematic experimental approach involving targeted saturation mutagenesis, selection of drug-resistant variants, and deep sequencing. We identified a single nucleotide substitution (T1514A, encoding L505H) that greatly increased drug resistance in cultured cells and mouse xenografts. The kinase activity of BRAFV600E/L505H was higher than that of BRAFV600E, resulting in cross-resistance to a MEK inhibitor. However, BRAFV600E/L505H was less resistant to several other BRAF inhibitors whose binding sites were further from L505 than that of PLX4720. Our results identify a novel vemurafenib-resistant mutant and provide insights into the treatment for melanomas bearing this mutation.
OI Bolon, Daniel/0000-0001-5857-6676
SN 1755-1471
EI 1755-148X
PD JAN
PY 2014
VL 27
IS 1
DI 10.1111/pcmr.12171
UT WOS:000328631100018
PM 24112705
ER

PT J
AU Hietpas, RT
Bank, C
Jensen, JD
Bolon, DNA

AF Hietpas, Ryan T.
Bank, Claudia
Jensen, Jeffrey D.
Bolon, Daniel N. A.
TI SHIFTING FITNESS LANDSCAPES IN RESPONSE TO ALTERED ENVIRONMENTS
SO EVOLUTION
AB The role of adaptation in molecular evolution has been contentious for decades. Here, we shed light on the adaptive potential in Saccharomyces cerevisiae by presenting systematic fitness measurements for all possible point mutations in a region of Hsp90 under four environmental conditions. Under elevated salinity, we observe numerous beneficial mutations with growth advantages up to 7% relative to the wild type. All of these beneficial mutations were observed to be associated with high costs of adaptation. We thus demonstrate that an essential protein can harbor adaptive potential upon an environmental challenge, and report a remarkable fit of the data to a version of Fisher's geometric model that focuses on the fitness trade-offs between mutations in different environments.
OI Bank, Claudia/0000-0003-4730-758X; Bolon, Daniel/0000-0001-5857-6676
SN 0014-3820
EI 1558-5646
PD DEC
PY 2013
VL 67
IS 12
BP 3512
EP 3522
DI 10.1111/evo.12207
UT WOS:000327572400011
PM 24299404
ER

PT J
AU Ryan, CJ
Cimermancic, P
Szpiech, ZA
Sali, A
Hernandez, RD
Krogan, NJ
AF Ryan, Colm J.
Cimermancic, Peter
Szpiech, Zachary A.
Sali, Andrej
Hernandez, Ryan D.
Krogan, Nevan J.
TI High-resolution network biology: connecting sequence with function
SO NATURE REVIEWS GENETICS
AB Proteins are not monolithic entities; rather, they can contain multiple domains that mediate distinct interactions, and their functionality can be regulated through post-translational modifications at multiple distinct sites. Traditionally, network biology has ignored such properties of proteins and has instead examined either the physical interactions of whole proteins or the consequences of removing entire genes. In this Review, we discuss experimental and computational methods to increase the resolution of protein-protein, genetic and drug-gene interaction studies to the domain and residue levels. Such work will be crucial for using interaction networks to connect sequence and structural information, and to understand the biological consequences of disease-associated mutations, which will hopefully lead to more effective therapeutic strategies.
OI Ryan, Colm/0000-0003-2750-9854
SN 1471-0056
EI 1471-0064
PD DEC
PY 2013
VL 14
IS 12
BP 865

EP 879
DI 10.1038/nrg3574
UT WOS:000327442800011
PM 24197012
ER

PT J
AU Harms, MJ
  Thornton, JW
AF Harms, Michael J.
  Thornton, Joseph W.
TI Evolutionary biochemistry: revealing the historical and physical causes
  of protein properties
SO NATURE REVIEWS GENETICS
AB The repertoire of proteins and nucleic acids in the living world is determined by evolution; their properties are determined by the laws of physics and chemistry. Explanations of these two kinds of causality - the purviews of evolutionary biology and biochemistry, respectively - are typically pursued in isolation, but many fundamental questions fall squarely at the interface of fields. Here we articulate the paradigm of evolutionary biochemistry, which aims to dissect the physical mechanisms and evolutionary processes by which biological molecules diversified and to reveal how their physical architecture facilitates and constrains their evolution. We show how an integration of evolution with biochemistry moves us towards a more complete understanding of why biological molecules have the properties that they do.
SN 1471-0056
PD AUG
PY 2013
VL 14
IS 8
BP 559
EP 571
DI 10.1038/nrg3540
UT WOS:000321956900011
PM 23864121
ER

PT J
AU Jiang, L
  Mishra, P
  Hietpas, RT
  Zeldovich, KB
  Bolon, DNA
AF Jiang, Li
  Mishra, Parul
  Hietpas, Ryan T.
  Zeldovich, Konstantin B.
  Bolon, Daniel N. A.
TI Latent Effects of Hsp90 Mutants Revealed at Reduced Expression Levels
SO PLOS GENETICS
AB In natural systems, selection acts on both protein sequence and expression level, but it is unclear how selection integrates over these two dimensions. We recently developed the EMPIRIC approach to systematically determine the fitness effects of all possible point mutants for important regions of essential genes in yeast. Here, we systematically investigated the fitness effects of point mutations in a putative substrate binding loop of yeast Hsp90 (Hsp82) over a broad range of expression strengths. Negative epistasis between reduced expression strength and amino acid substitutions was common, and the endogenous expression strength frequently obscured mutant defects. By analyzing fitness effects at varied expression strengths, we were able to uncover all mutant effects on function. The majority of mutants caused partial functional defects, consistent with this region of Hsp90 contributing to a mutation sensitive and critical process. These results demonstrate that important functional regions of proteins can tolerate mutational defects without experimentally observable impacts on fitness.
OI Bolon, Daniel/0000-0001-5857-6676

SN 1553-7404
PD JUN
PY 2013
VL 9
IS 6
AR e1003600
DI 10.1371/journal.pgen.1003600
UT WOS:000321222600066
PM 23825969
ER

PT J
AU Roscoe, BP
Thayer, KM
Zeldovich, KB
Fushman, D
Bolon, DNA
AF Roscoe, Benjamin P.
Thayer, Kelly M.
Zeldovich, Konstantin B.
Fushman, David
Bolon, Daniel N. A.
TI Analyses of the Effects of All Ubiquitin Point Mutants on Yeast Growth Rate
SO JOURNAL OF MOLECULAR BIOLOGY
AB The amino acid sequence of a protein governs its function. We used bulk competition and focused deep sequencing to investigate the effects of all ubiquitin point mutants on yeast growth rate. Many aspects of ubiquitin function have been carefully studied, which enabled interpretation of our growth analyses in light of a rich structural, biophysical and biochemical knowledge base. In one highly sensitive cluster on the surface of ubiquitin, almost every amino acid substitution caused growth defects. In contrast, the opposite face tolerated virtually all possible substitutions. Surface locations between these two faces exhibited intermediate mutational tolerance. The sensitive face corresponds to the known interface for many binding partners. Across all surface positions, we observe a strong correlation between burial at structurally characterized interfaces and the number of amino acid substitutions compatible with robust growth. This result indicates that binding is a dominant determinant of ubiquitin function. In the solvent-inaccessible core of ubiquitin, all positions tolerated a limited number of substitutions, with hydrophobic amino acids especially interchangeable. Some mutations null for yeast growth were previously shown to populate folded conformations indicating that, for these mutants, subtle changes to conformation caused functional defects. The most sensitive region to mutation within the core was located near the C-terminus that is a focal binding site for many critical binding partners. These results indicate that core mutations may frequently cause functional defects through subtle disturbances to structure or dynamics.
OI Bolon, Daniel/0000-0001-5857-6676
SN 0022-2836
EI 1089-8638
PD APR 26
PY 2013
VL 425
IS 8
BP 1363
EP 1377
DI 10.1016/j.jmb.2013.01.032
UT WOS:000317796200010
PM 23376099
ER

EF

Heat map vs mutational scan

Wednesday, September 28, 2016 11:59 PM

Made-up example:

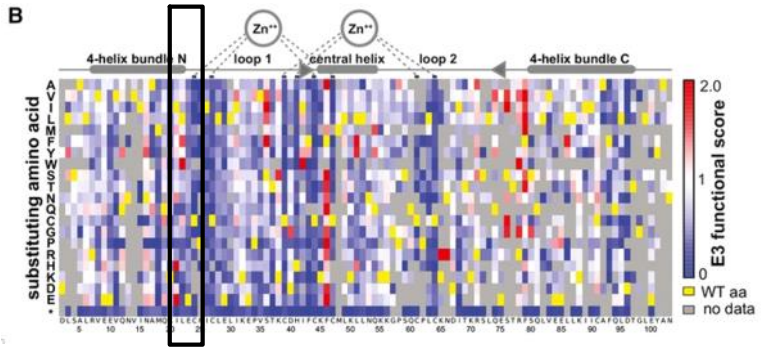
	A	R	T	M
W	0	0.3	0.3	0.3
F	0	0.6	0.9	0.6
Q	0.3	0.6	0.9	0.6
P	0.3	0.3	0.3	0.3

1 AA Sliding window

or

	1-25	26-50	51-75	76-100
1-25	0	0.3	0.3	0.3
26-50	0	0.6	0.9	0.6
51-75	0.3	0.6	0.9	0.6
76-100	0.3	0.3	0.3	0.3

25 AA Block window



	A	R	T	M
W	0	0.3	0.3	0.3
F	0	0.6	0.9	0.6
Q	0.3	0.6	0.9	0.6
P	0.3	0.3	0.3	0.3
Average	0.15	0.45	0.6	0.45

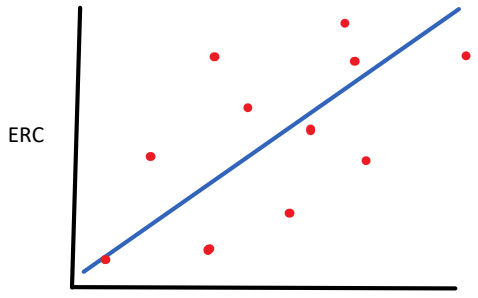
Might want deep mutational scanning (DMS) matrix for both members of the pair. Could compare ERC matrix with some generated pair-functional-score matrix. (Each cell an average of the particular amino acid scores?)

If only DMS matrix available, might be able to find some average of the ERC values in the single sequence present in the pair or sequences

Some average score for each amino acid as proxy for functional importance
May need to normalize somehow (some amino acid substitutions always more destructive than others)

*Variance instead of magnitude?

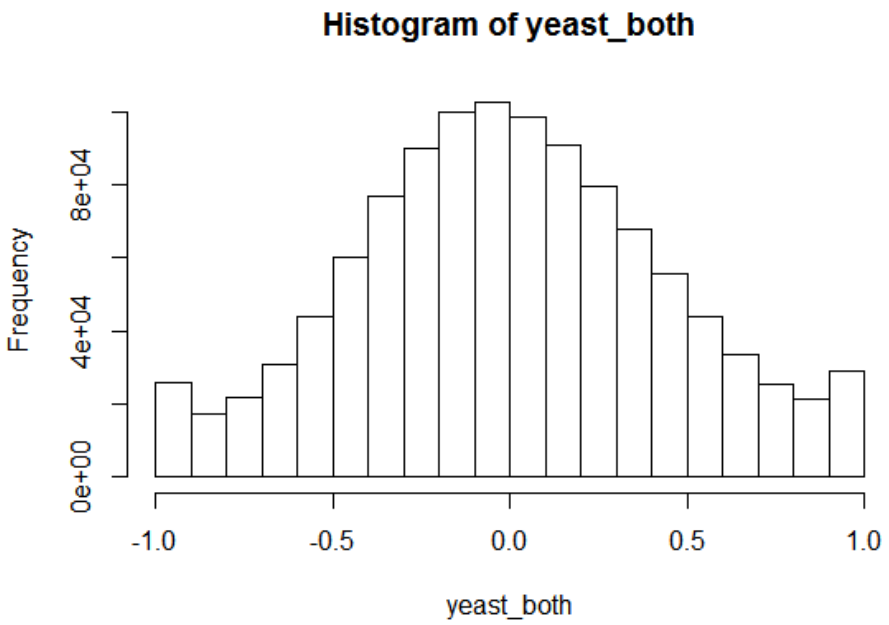
Plot ERC avg values vs functional scores



Heat map thoughts

Thursday, September 29, 2016 12:10 PM

Heat map windows compared to domain sequences for ERC script to produce average tree. Could be problematic?



Taking the average inflation of brians genes 65->1800
Applying it to all our data 5000->123,000
Means a 123,000 X 123,000 matrix for ERC script to analyze

Domain ERC matrix is ~10,000
This is 15,000,000,000 / 100,000,000 = 150 times larger, if , say ERC script takes 8 hours, 8 / 24 * 150 = 50 days. Would be no way unless ERC script broken down to parallel parts

Interpretation/Notes

Thursday, September 29, 2016 2:17 PM

What about finding sites that are evolutionarily conserved?
GERP (Genomic Evolutionary Rate Profiling) score measures evolutionary conservation of genetic sequences across species.

From <https://en.wikipedia.org/wiki/Conserved_sequence>

Genomic Evolutionary Rate Profiling: GERP

GERP identifies constrained elements in multiple alignments by quantifying substitution deficits. These deficits represent substitutions that would have occurred if the element were neutral DNA, but did not occur because the element has been under functional constraint. We refer to these deficits as "**Rejected Substitutions**". Rejected substitutions are a natural measure of constraint that reflects the strength of past purifying selection on the element. GERP estimates constraint for each alignment column; elements are identified as excess aggregations of constrained columns. A false-positive rate (which is user-settable) is calculated using 'shuffled' alignments in which the order of columns is randomized.

GERP++ Programs

GERP++ (previously referred to as GERP2) consists of two programs,*gerpcol* and *gerpelem*. Gerpcol esimates constraint for each column of the alignment; gerpelem then identifies constrained elements from gerpcol's output.

From <http://mendel.stanford.edu/SidowLab/downloads/gerp/>

*Change background to whole protein sequences instead of domains, correlate ERC values to see if difference.

10/6:

Null model: have outliers in heat map hot spot detection distribution (alyssa's). Proteins expressed in different places/times

Permute rows and columns (scramble) of erc

Email Ryan with

SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models

From <http://bioinformatics.oxfordjournals.org/content/28/20/2600.long>

Eugene wigner
Properties of random matrices
Energies of atomic nuclei

Meet w/ Alyssa

Friday, August 26, 2016 1:19 PM

Networkx

Analyze Heat map

Analyze Domain types

Vertebrates on HPC

Thursday, September 8, 2016 2:22 PM

- *Running parser --> will result in list of genes
- Muscle (just need to subdivide into more instances)
- Format alignments for PAML, set up files and folders
  (will need to subdivide more)
- *Run PAML on HPC, check PAML scores, make list of trees
- *Run ERC script in R on HPC

Also: should check Brian's domain ERC values with mine. Probably will have to re-run ERC script with just my domains for that info.

**Want to do unit tests on previous steps to have more formal way of ensuring accuracy

Hpc.arizona.edu

Check for PAML documentation on sequence sizes vs number species

Vertibrates on HPC

Wednesday, September 14, 2016 10:38 PM

Waiting for:
- sap-por_par
- sap-pri_par

Ask ryan about specifc papers on deep mutational scanning he has in mind

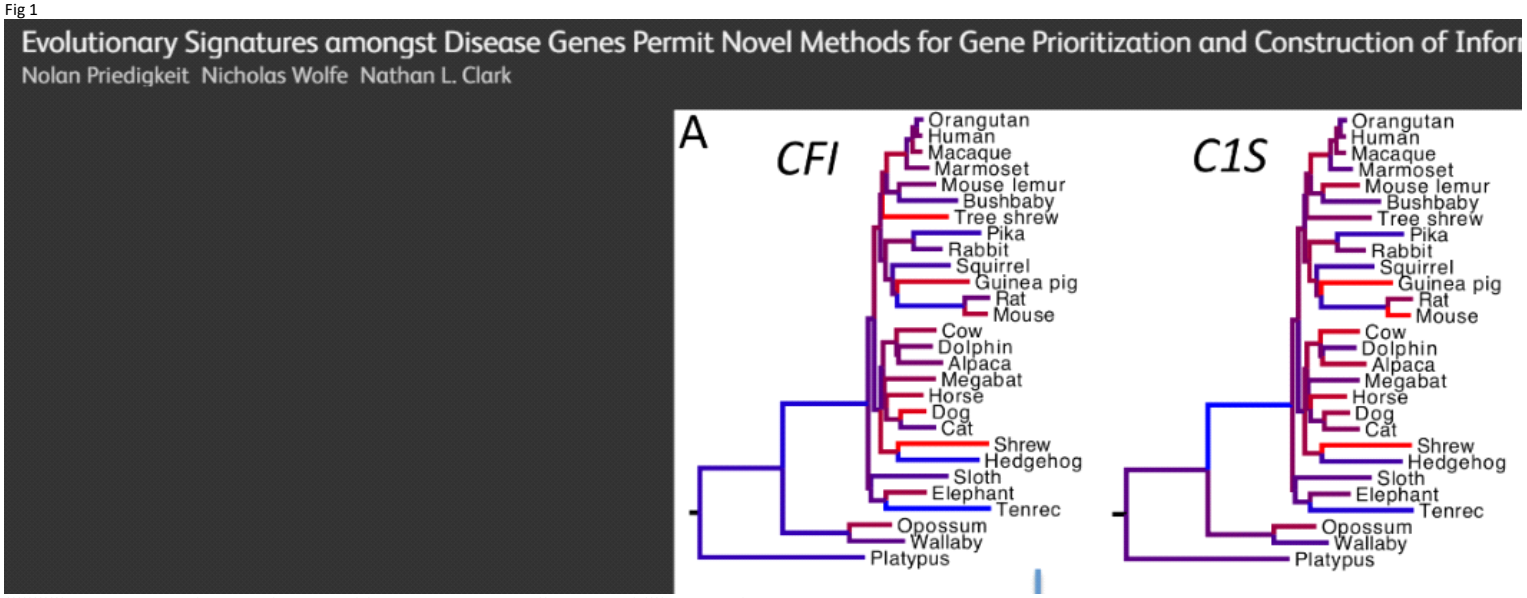
-added 2 species
-muscle (htc vs ocelot)
-phylogenetic tree not sited
-paml
-erc

Phylogenetic tree

Tuesday, September 20, 2016 2:35 PM

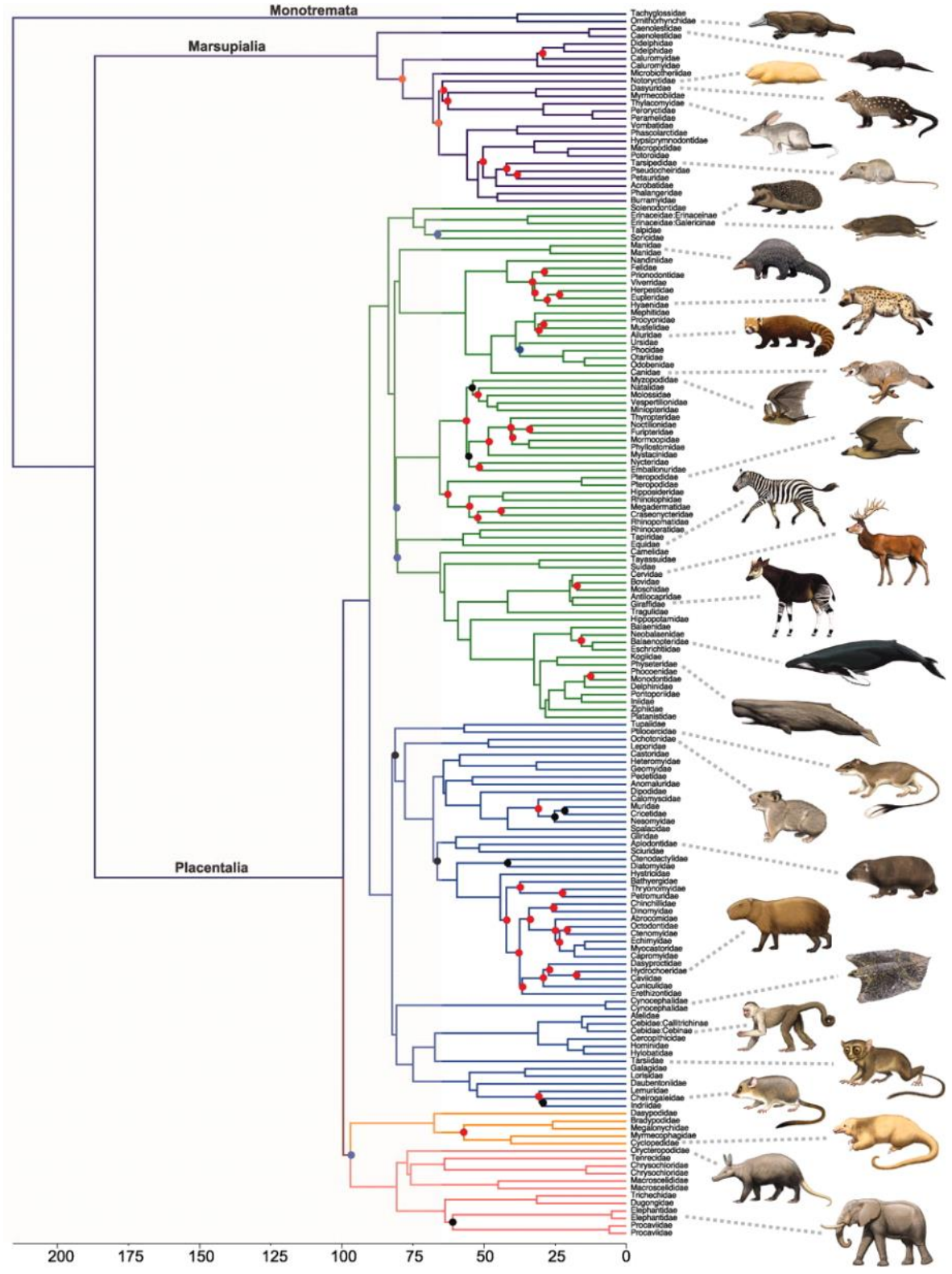
http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004967

Fig 1



((((((((((Orangutan,Human),Macaque),Marmoset),(Mouse lemur,Bushbaby)),Tree shrew),((((Rat,Mouse),Guinea pig),Squirrel),(Rabbit,Pika))),(((((Dog,Cat),Horse),Megabat),((Dolphin,Alpaca),Cow)),(Shrew,Hedgehog))),((Elephant,Tenrec),Sloth)),(Opossum,Wallaby)),Platypus)

http://science.sciencemag.org/content/334/6055/458.full



http://www.timetree.org/book
http://www.hedgeslab.org/pubs/260.pdf

Tree

Wednesday, September 21, 2016 9:29 PM

include Fish?

Brian:
~~Gallus gallus~~, Pan troglodytes, Danio rerio, Mus musculus, Rattus norvegicus, Bos taurus, Canis lupus familiaris, Homo sapiens

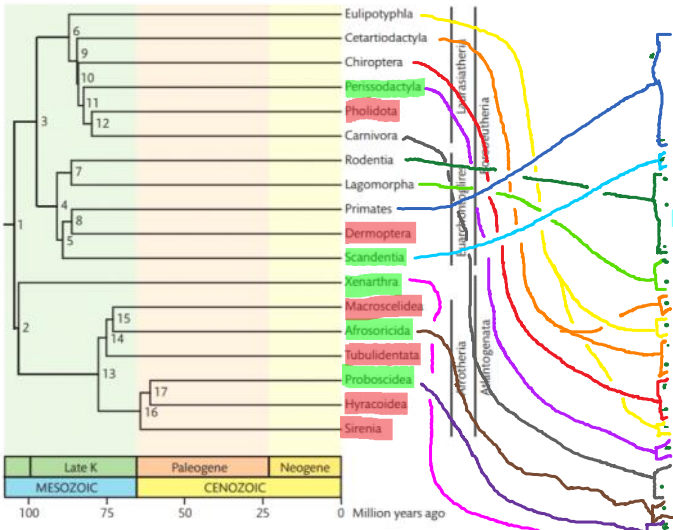
Nathan (https://files.acrobat.com/a/preview/f2b9388e-75bd-45f6-a021-855fe0fcd1ee)
Homo sapiens (human), Pongo pygmaeus abelii (orangutan), Macaca mulatta (rhesus macaque), Callithrix jacchus (marmoset), ~~Tarsius syrichta~~ (tarsier), Microcebus murinus (mouse lemur), Otolemur garnettii (bushbaby), Tupaia belangeri (tree shrew), Cavia porcellus (guinea pig), Dipodomys ordii (kangaroo rat), Mus musculus (mouse), Rattus norvegicus (rat), ~~Spermophilus tridecemlineatus~~ (squirrel), Oryctolagus cuniculus (rabbit), Ochotona princeps (pika), Vicugna pacos (alpaca), Sorex araneus (shrew), Bos taurus (cow), Tursiops truncatus (dolphin), Pteropus vampyrus (megabat), Myotis lucifugus (microbat), Erinaceus europaeus (hedgehog), Equus caballus (horse), Canis lupus familiaris (dog), Felis catus (cat), ~~Choloepus hoffmanni~~ (sloth), Echinops telfairi (tenrec), Loxodonta africana (elephant), ~~Procavia capensis~~ (rock hyrax), Dasypus novemcinctus (armadillo), Monodelphis domestica (opossum), ~~Macropus eugenii~~ (wallaby), and Ornithorhynchus anatinus (platypus)

Shared species ~~Not in refseq~~

Question Found in refseq

30 of these species found in refseq

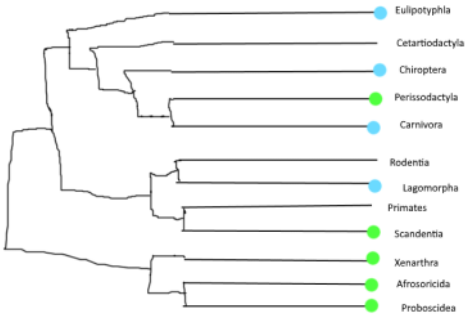
Placental mammals (Eutheria)
http://www.timetree.org/public/data/pdf/Murphy2009Chap71.pdf



Completed inparanoid runs:

	Order:	Family:
~~Pan troglodytes~~(chimp)	Primates	Hominidae
Pongo pygmaeus abelii (orangutan),	Primates	Hominidae
Macaca mulatta (rhesus macaque),	Primates	Cercopithecidae
Callithrix jacchus (marmoset),	Primates	Cedbidae
Microcebus murinus (mouse lemur),	Primates	Cheirogaleidae
Otolemur garnettii (bushbaby),	Primates	Galagonidae
Tupaia belangeri (tree shrew),	**Scandentia**	
Cavia porcellus (guinea pig),	Rodentia	Caviidae
Dipodomys ordii (kangaroo rat),	Rodentia	Heteromyidae
Mus musculus (mouse),	Rodentia	Muridae
Rattus norvegicus (rat),	Rodentia	Muridae
Oryctolagus cuniculus (rabbit),	Lagomorpha	Leporidae
Ochotona princeps (pika),	Lagomorpha	Ochotonidae
Vicugna pacos (alpaca),	Cetariodactyla	Camelidae
Sorex araneus (shrew),	Eulipotyphla	Sorcidae
Bos taurus(cow),	Cetariodactyla	Bovidae
Tursiops truncatus (dolphin),	Cetariodactyla	Delphinidae
Pteropus vampyrus (megabat),	Chiroptera	Pteropodidae
Myotis lucifugus (microbat),	Chiroptera	Vespertilionidae
Erinaceus europaeus (hedgehog),	Eulipotyphla	Erinaceidae
Equus caballus (horse),	Perissodactyla	
Canis lupus familiaris (dog),	Carnivora	Canidae
Felis catus (cat),	Carnivora	Felidae
Echinops telfairi (tenrec),	Afrosoricida	
Loxodonta africana(elephant),	Proboscidea	
Dasypus novemcinctus (armadillo),	Xenartha	
~~Monodelphis domestica~~ (opossum),	Didelphimorphia	
Ornithorhynchus anatinus (platypus)	Monotremata	

Exit status = 0
Exit status = 271 (Possibly ran out of memory)
Present in phylo tree in Nathan's paper
Still running (9/1)
Have output for [18/28]

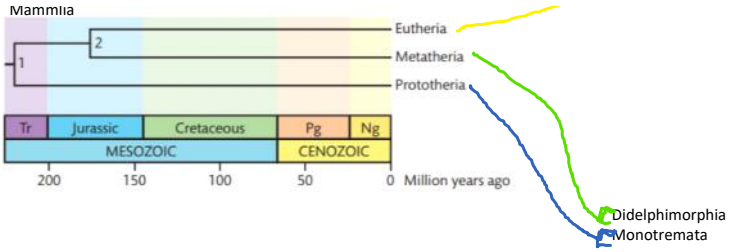


((((((Carnivora,Perissodactyla),Chiroptera),Cetartiodactyla),Eulipotyphla),((Rodentia,Lagomorpha),(Primates,Scandentia))),((Afrosoricida,Proboscidea),Xenarthra))

(((((((Canidae,Felidae),Perissodactyla),(Pteropodidae,Vespertilionidae)),((Delphinidae,Bovidae),Camelidae)),(Sorcidae,Erinaceidae)),((((Muridae,Heteromyidae),Caviidae),(Leporidae,Ochotonidae)),((((Hominidae,Cercopithecidae),Cebidae),(Galagonidae,Cheriogaleidae)),Scandentia))),((Afrosoricida,Proboscidea),Xenarthra))

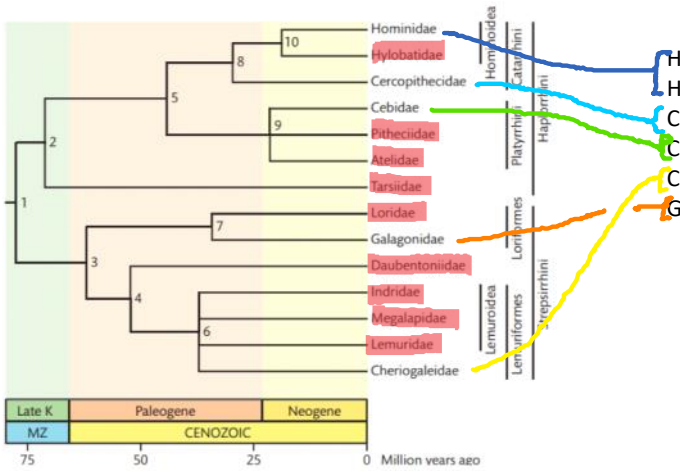
((((((((((familiari,catus),caballus),(vampyrus,lucifugus)),((truncatus,taurus),pacos)),(araneus,europaeus)),(((((norvegicu,musculus),ordii),porcellus),(cuniculus,princeps)),((((((sapiens,troglodyt),abelii),mulatta),jacchus),(garnettii,murinus)),belangeri))),((telfairi,africana),novemcinc)),domestica),anatinus);

http://www.timetree.org/public/data/pdf/Madsen2009Chap68.pdf
Mammlia

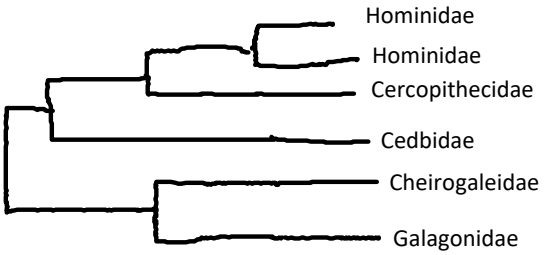

Mammlia
Eutheria
Metatheria
Prototheria
2
1
Tr
Jurassic
Cretaceous
Pg
Ng
MESOZOIC
CENOZOIC
200
150
100
50
0
Million years ago
Didelphimorphia
Monotremata

*Primates

Wednesday, September 21, 2016 10:51 PM



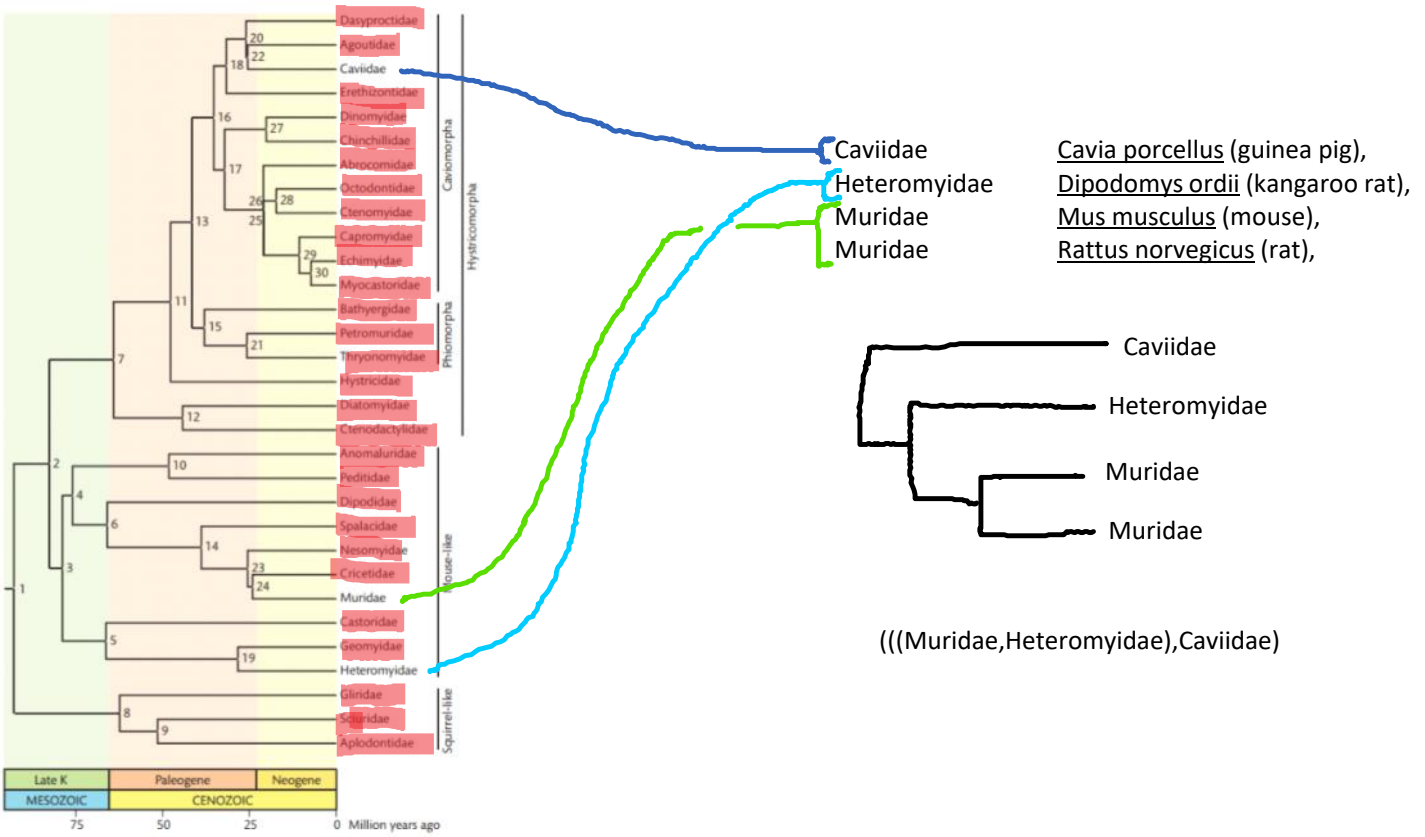
Hominidae	Primates	Pan troglodytes(chimp)
Hominidae	Primates	Pongo pygmaeus abelii (orangutan),
Cercopithecidae	Primates	Macaca mulatta (rhesus macaque),
Cebidae	Primates	Callithrix jacchus (marmoset),
Cheirogaleidae	Primates	Microcebus murinus (mouse lemur),
Galagonidae	Primates	Otolemur garnettii (bushbaby),



(((Hominidae, Cercopithecidae),Cebidae),(Galagonidae,Cheriogaleidae))

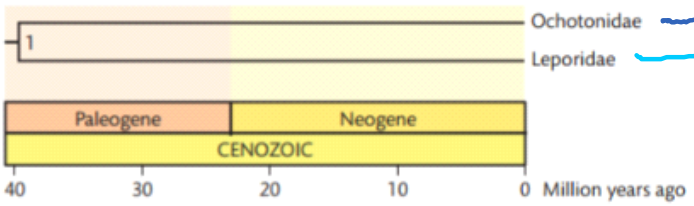
*Rodentia

Wednesday, September 21, 2016 10:50 PM



Lagomorpha

Wednesday, September 21, 2016 10:53 PM



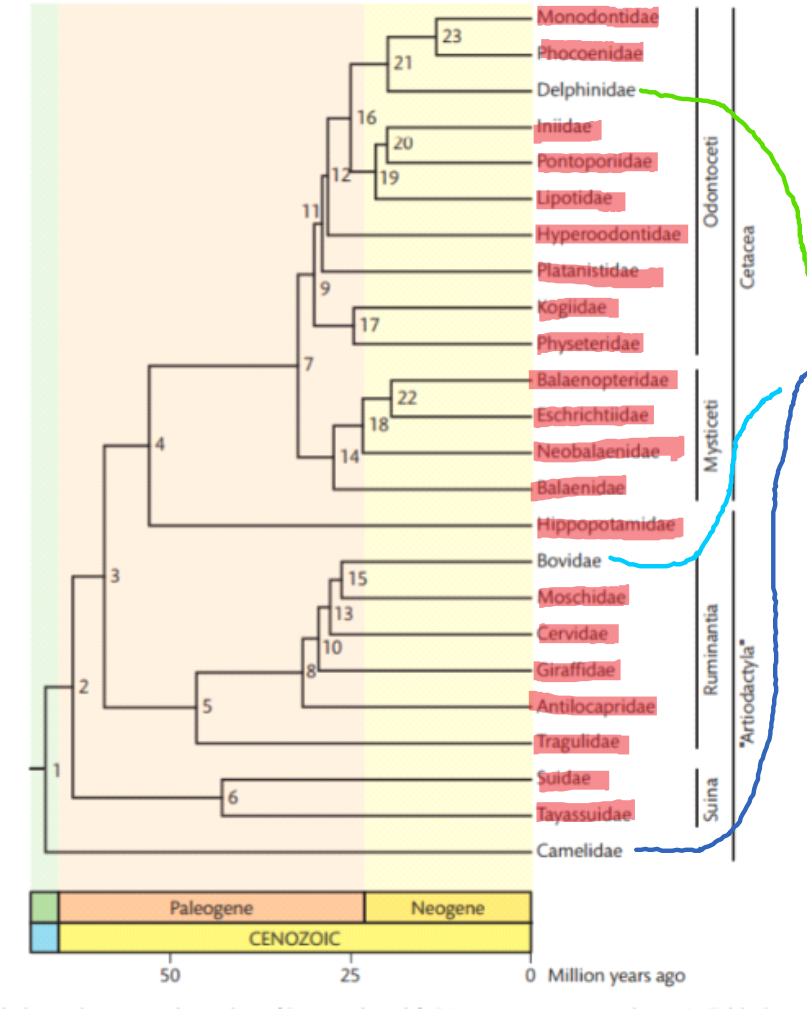
Leporidae
Ochotonidae

Oryctolagus cuniculus (rabbit),
Ochotona princeps (pika),

(Leporidae,Ochotonidae)

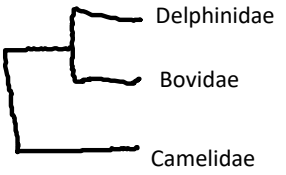
*Cetariodactyla

Wednesday, September 21, 2016 10:55 PM



Camelidae
Bovidae
Delphinidae

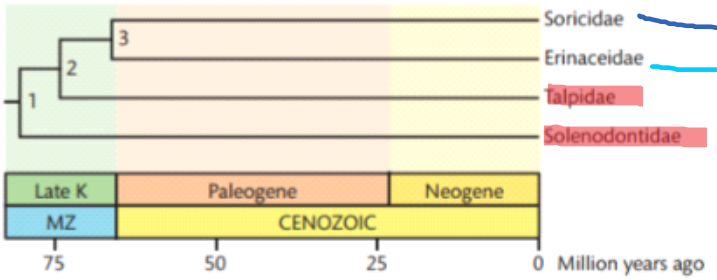
Vicugna pacos (alpaca),
Bos taurus(cow),
Tursiops truncatus (dolphin),



((Delphinidae,Bovidae),Camelidae)

Eulipotyphla

Wednesday, September 21, 2016 11:00 PM



Sorcidae
Erinaceidae

Sorex araneus (shrew),
Erinaceus europaeus (hedgehog),

(Sorcidae,Erinaceidae)

Chiroptera

Wednesday, September 21, 2016 11:02 PM

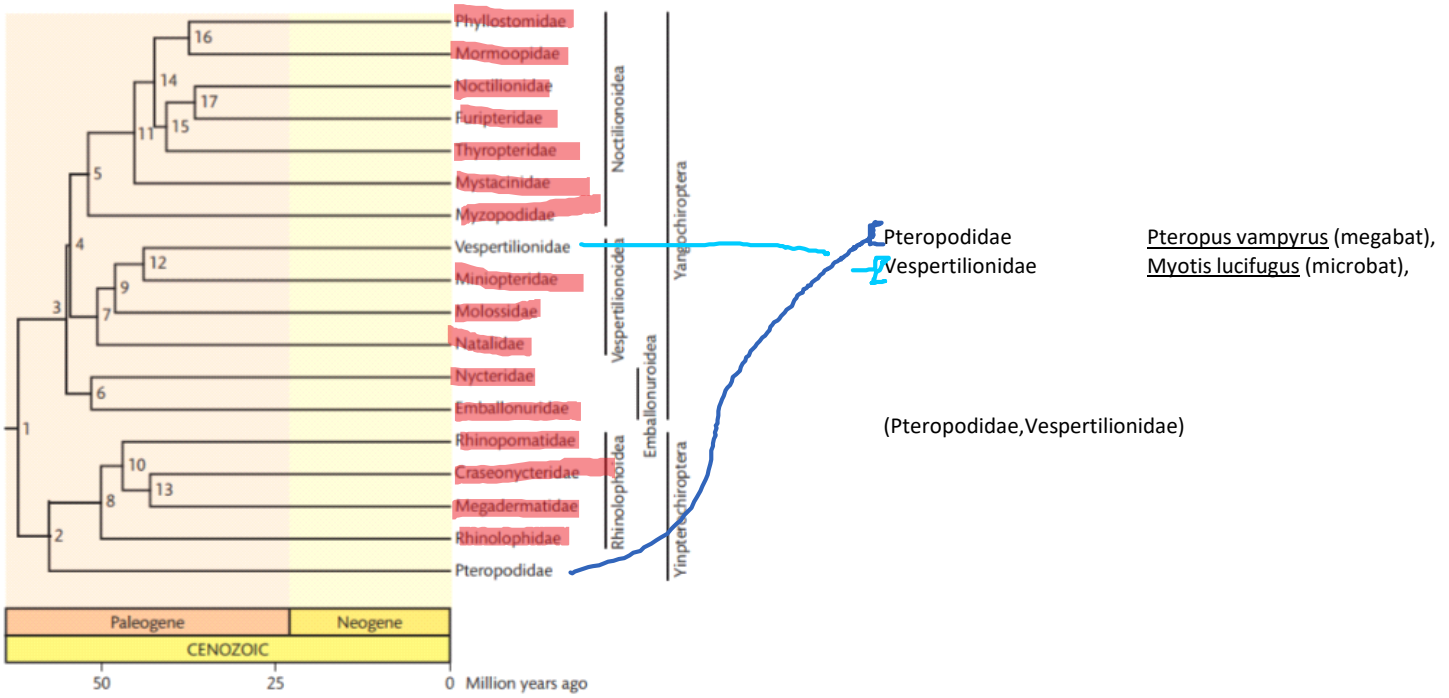
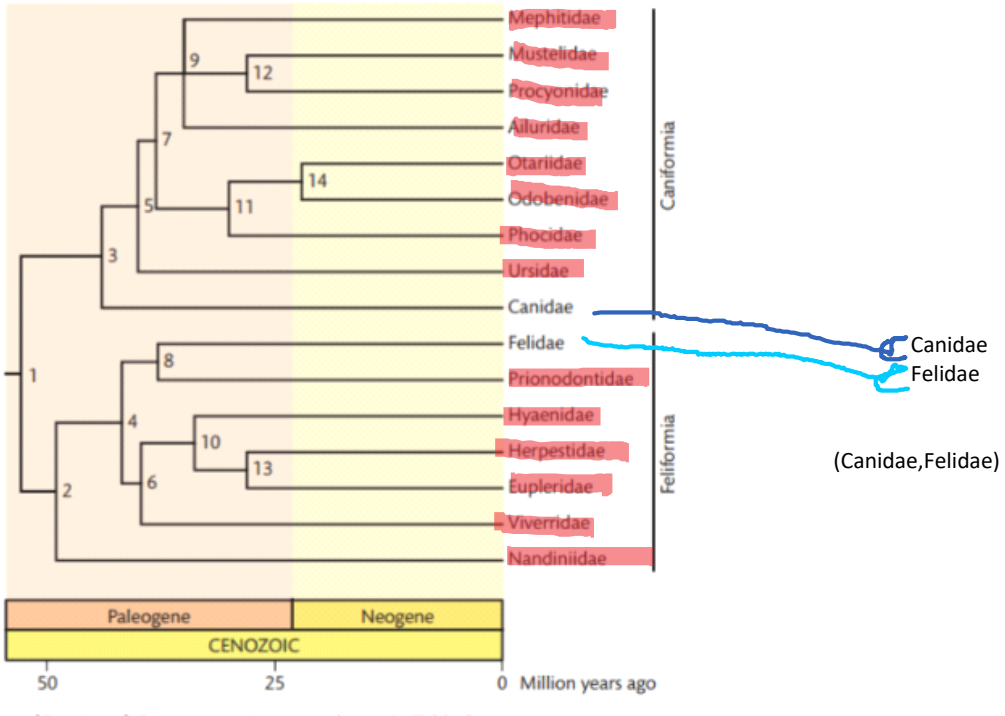


Fig. 2 A timetree of bats (Chiroptera). Divergence times are shown in Table 1.

Carnivora

Wednesday, September 21, 2016 11:05 PM



Canidae
Felidae

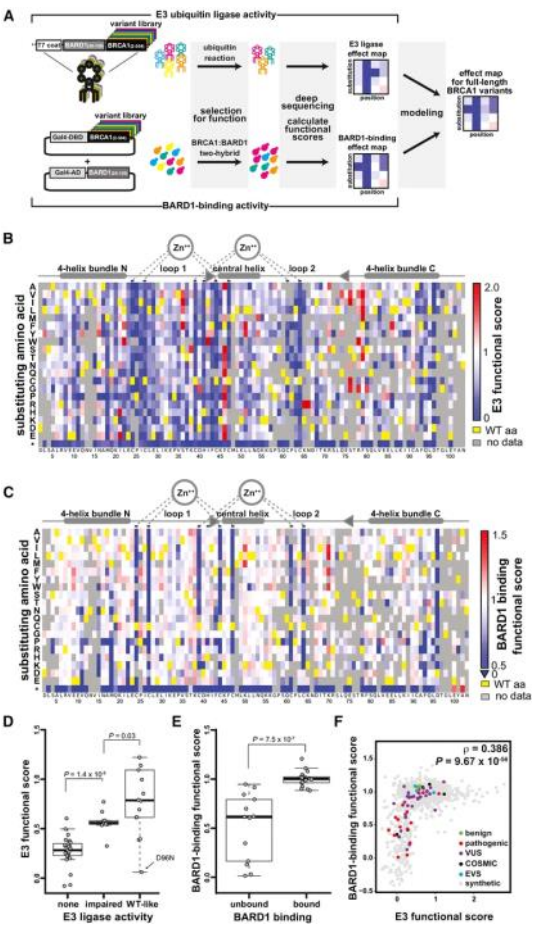
(Canidae,Felidae)

Canis lupus familiaris (dog),
Felis catus (cat),

Deep Mutational Scanning

Thursday, September 22, 2016 2:48 PM

http://www.genetics.org/content/200/2/413



(A) Scheme for leveraging scores from parallelized assays for BRCA1 RING function into predictions for the function of the full-length BRCA1 protein in homology-directed DNA repair. (B-F) Scoring the E3 ligase and BARD1-binding activities of BRCA1 RING domain variants. (B) A sequence-function map of the effect of missense mutations in the BRCA1 RING domain on E3 ligase function. The functional score for each variant is the slope of the fit curve, normalized by setting stop codons to a score of 0 and the wild-type to a score of 1. Each position in BRCA1(2-103) is arranged along the x-axes, structural features of the RING domain are diagrammed above. The amino acid substitutions, grouped by side-chain properties, are on the y-axes. The E3 ligase scores range from improved activity versus wild-type (red), equivalent to wild-type (white), to less than wild-type (blue). Yellow represents the wild-type residue and gray missing or low confidence data. (C) A sequence-function map of the effect of missense mutations in the BRCA1 RING domain on BARD1-RING binding. Coloring as in panel B. (D) Comparison of the variant scores from the deep mutational scan for E3 ligase activity versus literature-reported E3 ligase activities for the same BRCA1 variants (Brzovic *et al.*, 2003; Morris *et al.*, 2006). The Wilcoxon rank sum test (WRST) was used to test for significant differences between the categories. The biggest outlier in the wild type-like category, D96N, not only performed poorly as an E3 ligase score but also failed to bind to BARD1 and to support homology-directed repair in cells (Table S2). (E) Comparison of BARD1-binding scores from the two-hybrid experiment versus literature-reported BARD1 binding by the same BRCA1 variants (Brzovic *et al.*, 2003; Ransburgh *et al.*, 2010). The WRST was used to test for significant differences between categories. (F) The relationship between the quality-filtered E3 ligase functional scores and the BARD1-binding scores. Colors indicate the clinical classification or database of origin for each variant.

From <http://www.genetics.org/content/200/2/413>

Interpretation

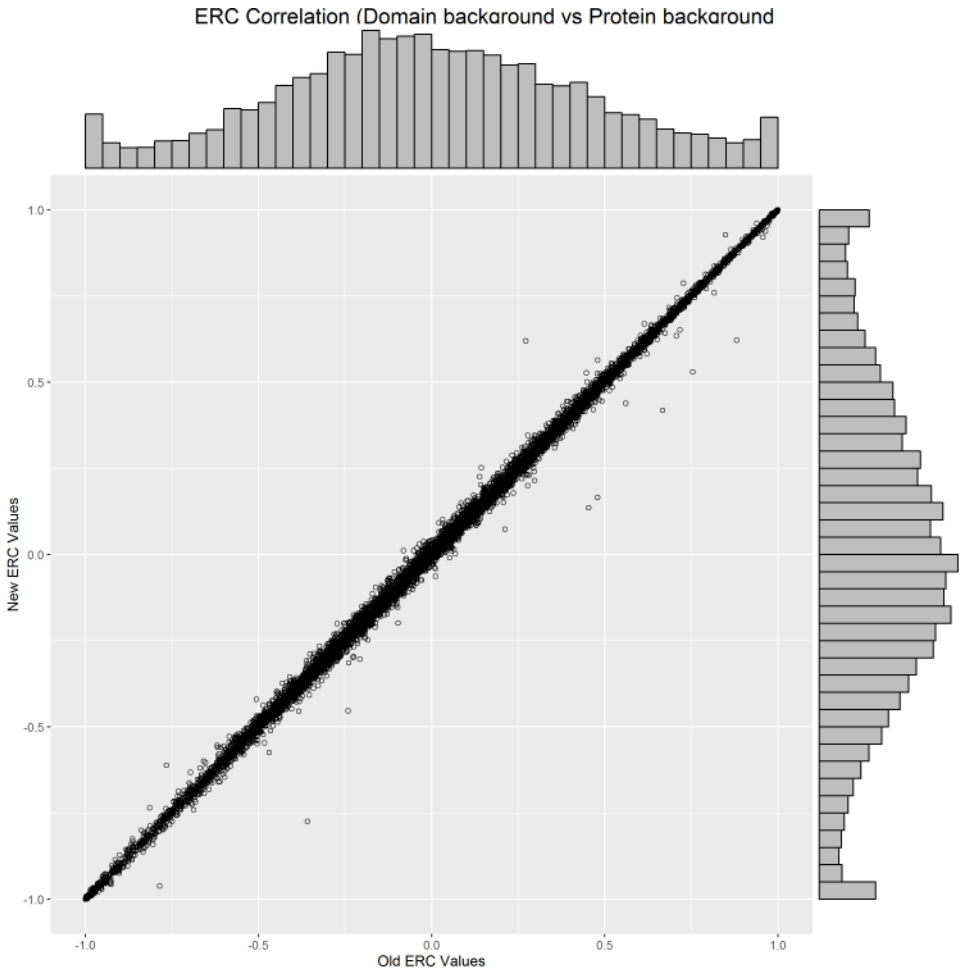
Thursday, September 22, 2016 3:28 PM

Check values
Look at deep mutational scanning, other examples
Copasi, help ryan with models, starts october 5th

SVD: singular value decomposition

Check background

Thursday, October 13, 2016 11:11 AM



Old: Domain background to window slices of Brian's proteins
New: Full protein background to window slices of Brian's proteins

Next steps

Thursday, October 13, 2016 11:14 AM

ERC vertibrates:
1st day: 25%
3rd day: 50%
5th day: 70%
7th day: 82%
8th day: 83%
15th day: 89%

Possible next steps:

- Network dynamical influence
- Heat maps
    - Find Brian's vertebrate proteins
        - Use window slicer on them
        - Translate:
            - http://www.ensembl.org/biomart/martview/bdf6f0be101401d0902de0a2ce6ac6aa
    - Run Brian's vertebrate proteins with full protein background with duplicates removed
        - Since full protein background pretty much equivalent, should run everything with it so we don't have to wait for the domain sliced (larger!) vertebrate data to come--
        - 2.5*60,000=150,000 things to run if vert sliced as domains, would take much longer
    - Pull out submatrix of Brian's genes
    - Include alignments in folder

Brian vert list

Thursday, October 13, 2016 2:03 PM

10532
1257
1645
1955
2119
22876
2999
3687
37912
44935
55429
55899
715
7738
8509
1057
128851
1673
20084
2122
23194
31000
37455
37990
455
55532
560
7242
77719
901
1065
13223
1678
20090
21249
237
31024
37525
3802
474
55549
561
73902
7813
90937
111276
1434
173

20457
21250
2614
3197
37637
380
48145
55562
68068
74409
7889
923
11819
14452
1855
20474
2163
265
3254
37643
3828
4846
55617
68203
74545
7934
9563
121574
1483
1876
2063
2168
272
3273
37659
4049
4fold
55621
6820
7497
7960
1222
1510
1898
20926
2177
2824
3454
37660
4117
502
55629
68719

7626
8135
12267
1542
19127
21084
2197
2845
35136
37670
426
50501
55679
68986
7652
81850
123904
1576
1954
21120
22398
2941
3637
3785
4299
527
55703
6945
7657
8359

Interpretation

Thursday, October 13, 2016 3:39 PM

Look at R script parallelizing, decrease memory allocation (memory use).

May want to re-run yeast heat-map slice data with protein background if we go further with it because there are average 2.5 domains(including linkers) per sequence, so running vertebrates with domain backgrounds will take a very long time.

Fix algorithm

Papers that use evolutionary data to find area of interaction
Are we solving the problem
Is it a question other people haven't already solved
(Comparing phylogeny and sequences of proteins, binding sites)

Parallel ERC

Thursday, November 3, 2016 2:43 PM

Preprocess data by filling in ERC matrix with an assigned topology number

A set of topologies assigned to each parallel slave

Runs through entire matrix, each slave has a subset of the matrix filled in

Master code combines each matrix subset into completed ERC matrix

Having issues getting parallel to run properly. Not always recognizing variables, apparently not running all code within slave process. Not returning what I want. Errors about "Subscript out of bounds"

Array job: environment variable (1-100 for example)
google pbs array jobs
Travis script
System decide when to queue (assign 1 CPUs)
qstat -f (resources not available, etc.)

Difficult to communicate between nodes, communication speed

Have script to assign job matrix

Need script for each job to just run its sections of the matrix. Where ii == true or something to make list of pairs of matrix indices to go through. So should only need to set for loop conditions. Everything else inside for loop should be able to remain the same. Submatrices get saved into special folder. Will have ERC and P value matrices. Will also need to pull in command line arguments.

Need script to put together sub matrices into complete matrix. Should just be able to iterate through them all to complete matrix.

Have script to call an array of jobs into R

Test script on yeast values to make sure they match

Run script on vertibrates

Memory issues: 47 gb ram per run

Can list genes containing topology along with topology ID #

Don't see any obvious reasons for huge memory usage yet

2-13-2017 Protein Domains

Monday, February 13, 2017 1:32 PM

ERC within the same protein.
Remove domains within the same protein from analysis.
Standard error? And mean --> p values, sort out interesting ones
Physical interactions , map of protein -protein interactions: look at ID2 stuff we did before
Q-Q plot, 2 distributions, how different. Compare quantiles etc. between distributions
Outline due next friday, look for form
Get done exploratory analysis

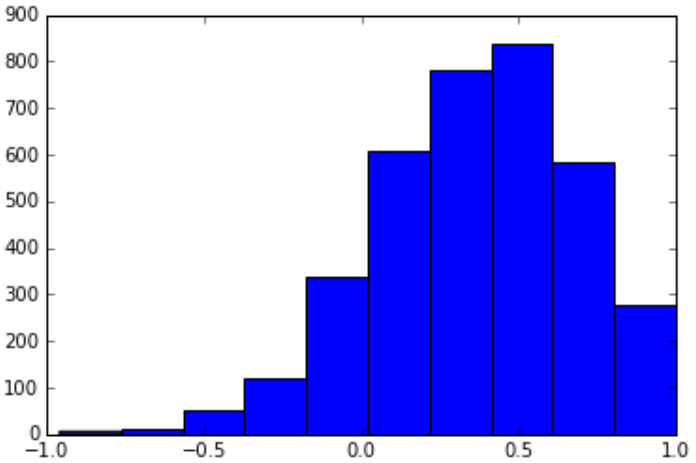
Motivation dynamical influence
Calc domain level erc, exploratory analysis, dynamic evolution

Compare interacting proteins and all others q-q plot
Definition of interacting (databases used by i2d) why use this definition, alternatives.

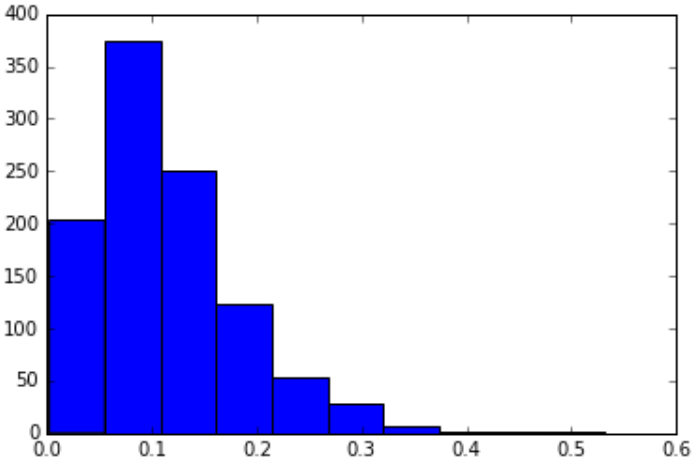
ERC values based on domains

Monday, February 27, 2017 12:12 PM

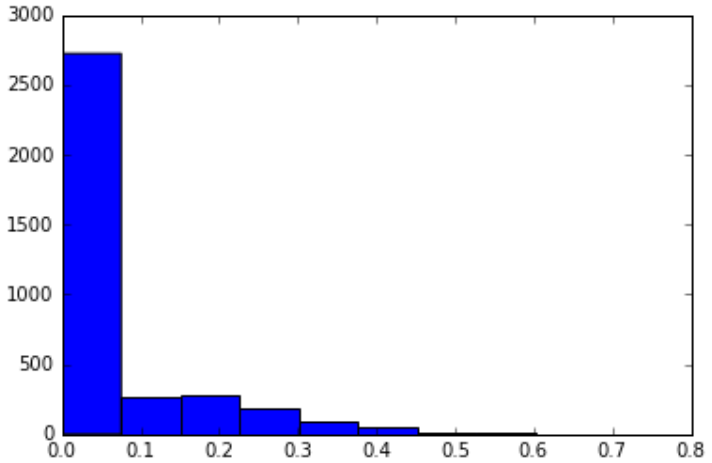
Between domains on the same protein:



Between domains on the same protein, mean is 0.362

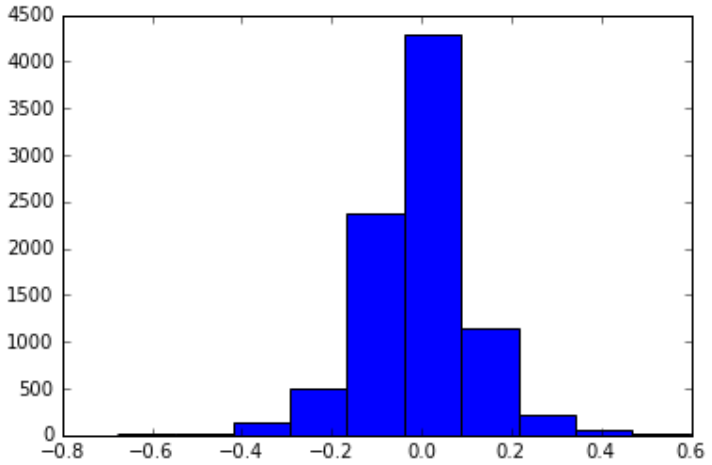


Standard error of the mean is 0.113

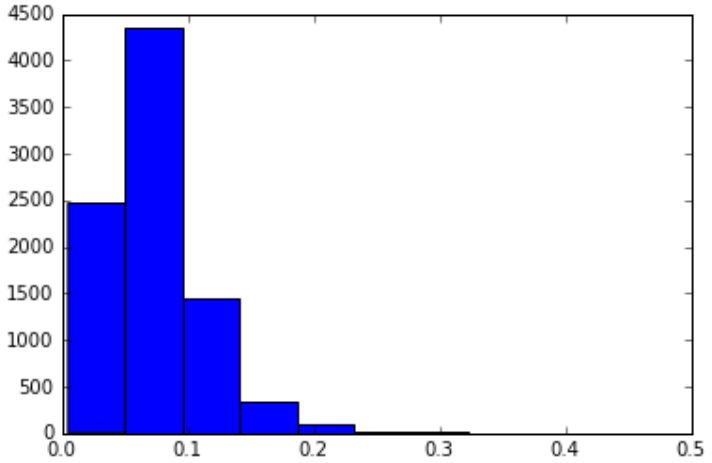


And std dev is 0.0549

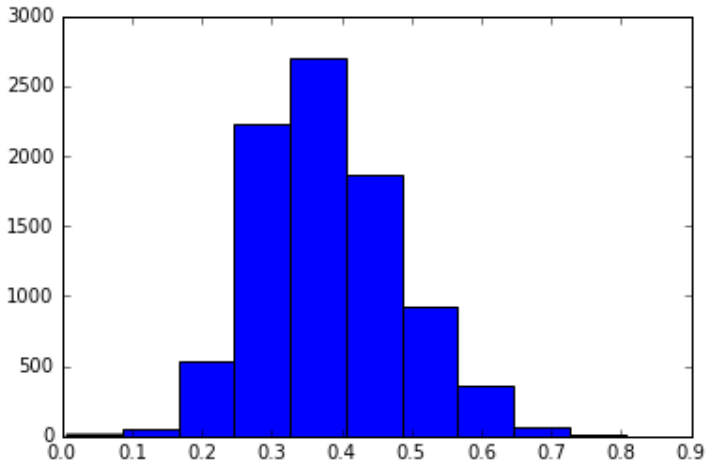
Between domain types in general:



Distribution of means between domain types (using average of means between each pair of domain types)
mean = -0.00589



Standard error of the mean = 0.0736



Std dev = 0.381

High ERC, low SEM, low SD

Monday, March 6, 2017 11:56 PM

('ABC2_membrane', 'FAD_binding_8')
['erc mean: 0.53', 'erc sem: 0.20', 'erc std dev: 0.40']

('Aldo_ket_red', 'Ferric_reduct')
['erc mean: 0.53', 'erc sem: 0.09', 'erc std dev: 0.27']

('Glyco_transf_15', 'Ferric_reduct')
['erc mean: 0.51', 'erc sem: 0.11', 'erc std dev: 0.35']

('HMG_box', 'FAD_binding_8')
['erc mean: 0.52', 'erc sem: 0.16', 'erc std dev: 0.50']

('NAD_binding_6', 'FAD_binding_8')
['erc mean: 0.59', 'erc sem: 0.15', 'erc std dev: 0.54']

('NAD_binding_6', 'Ferric_reduct')
['erc mean: 0.50', 'erc sem: 0.10', 'erc std dev: 0.43']

('NAD_binding_6', 'NAD_binding_6')
['erc mean: 0.54', 'erc sem: 0.17', 'erc std dev: 0.47']

('Snf7', 'Arf')
['erc mean: 0.60', 'erc sem: 0.04', 'erc std dev: 0.13']

('Snf7', 'Snf7')
['erc mean: 0.51', 'erc sem: 0.04', 'erc std dev: 0.17']

High ERC, low SEM, low SD. With pvals

Thursday, March 23, 2017 2:07 PM

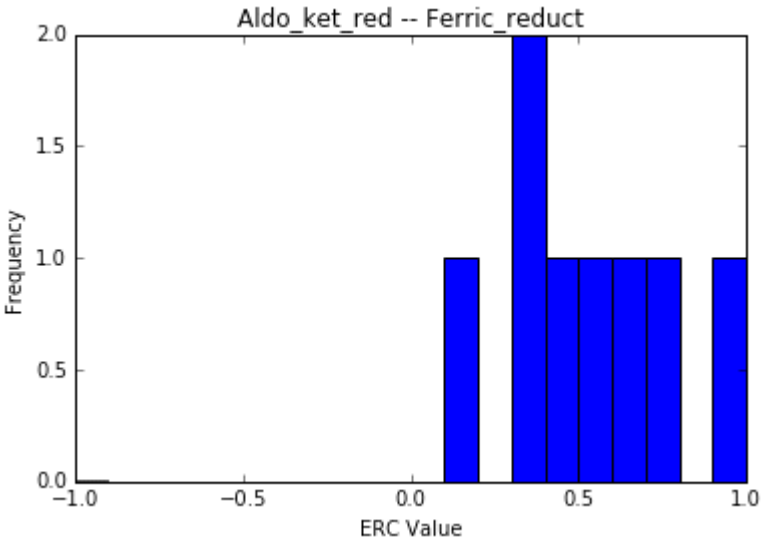
{('ABC2_membrane', 'FAD_binding_8'): ['erc mean: 0.5296',

'erc sem: 0.2014',

'erc std dev: 0.4028',

'erc tstat: 2.6298',

'erc pval: 0.0783'],



('Aldo_ket_red', 'Ferric_reduct'): ['erc mean: 0.5297',

'erc sem: 0.0938',

'erc std dev: 0.2652',

'erc tstat: 5.6502',

'erc pval: 0.0008'],

('Aldo_ket_red', 'Glyco_transf_15'): ['erc mean: 0.5284',

'erc sem: 0.0953',

'erc std dev: 0.3436',

'erc tstat: 5.5451',

'erc pval: 0.0001'],

('Cation_ATPase_N', 'adh_short'): ['erc mean: 0.5384',

'erc sem: 0.1115',

'erc std dev: 0.3862',

'erc tstat: 4.8295',

'erc pval: 0.0005'],

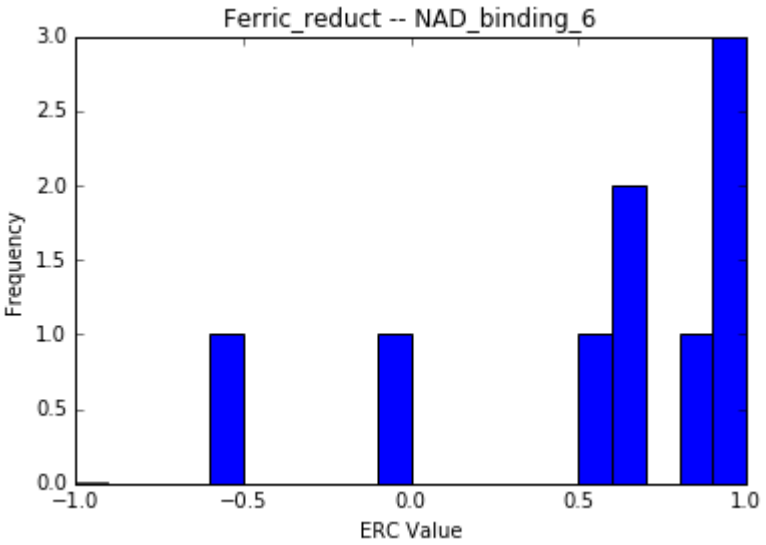
('Ferric_reduct', 'FAD_binding_8'): ['erc mean: 0.5138',

'erc sem: 0.1435',

'erc std dev: 0.5558',

'erc tstat: 3.5805',

'erc pval: 0.0030'],



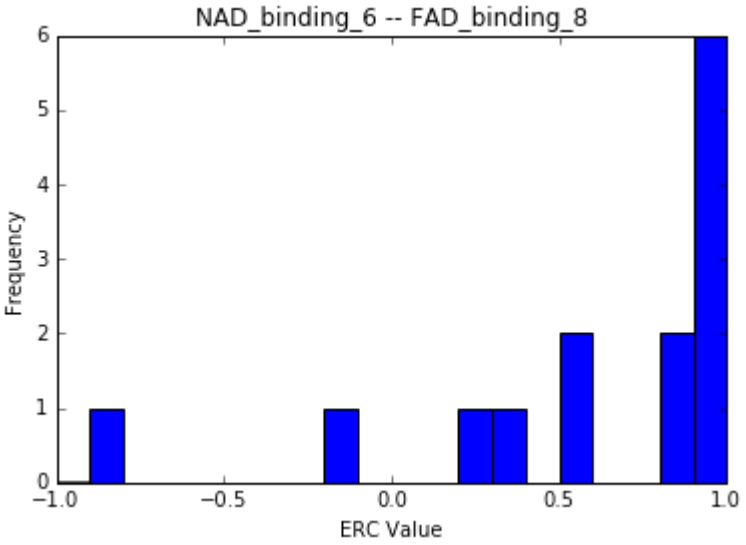
('Ferric_reduct', 'NAD_binding_6'): ['erc mean: 0.5427',

'erc sem: 0.1732',

'erc std dev: 0.5196',

'erc tstat: 3.1335',

'erc pval: 0.0139'],



('NAD_binding_6', 'FAD_binding_8'): ['erc mean: 0.5860',

'erc sem: 0.1456',

'erc std dev: 0.5447',

'erc tstat: 4.0253',

'erc pval: 0.0014'],

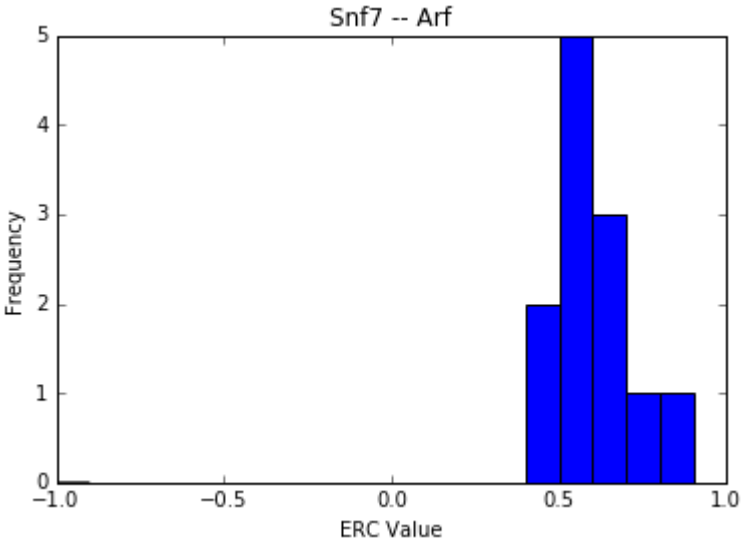
('NAD_binding_6', 'NAD_binding_6'): ['erc mean: 0.5358',

'erc sem: 0.1667',

'erc std dev: 0.4714',

'erc tstat: 3.2154',

'erc pval: 0.0147'],



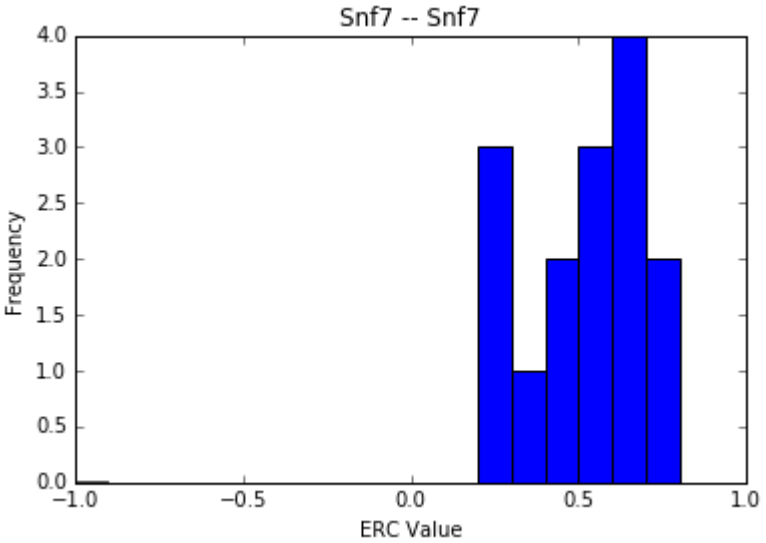
('Snf7', 'Arf'): ['erc mean: 0.5967',

'erc sem: 0.0368',

'erc std dev: 0.1276',

'erc tstat: 16.2052',

'erc pval: 0.0000'],



('Snf7', 'Snf7'): ['erc mean: 0.5097',

('Snf7', 'Snf7'): ['erc mean: 0.5097',

'erc sem: 0.0428',

'erc std dev: 0.1659',

'erc tstat: 11.8956',

'erc pval: 0.0000'],

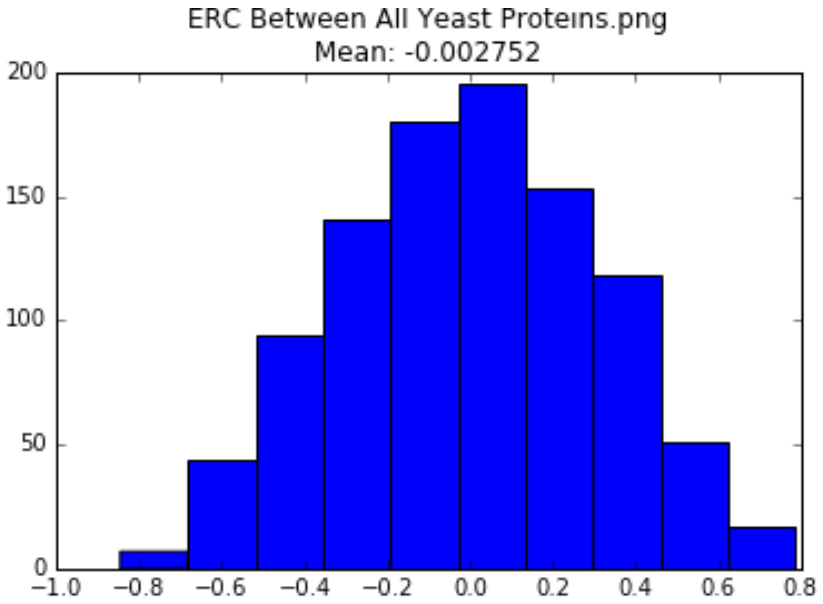
('Transp_cyt_pur', 'dCMP_cyt_deam_1'): ['erc mean: 0.5299',

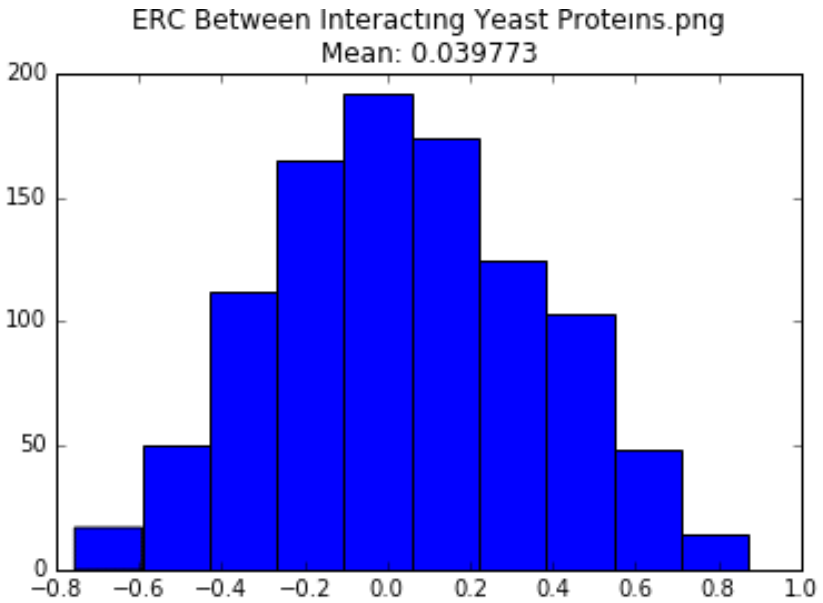
'erc sem: 0.1600',

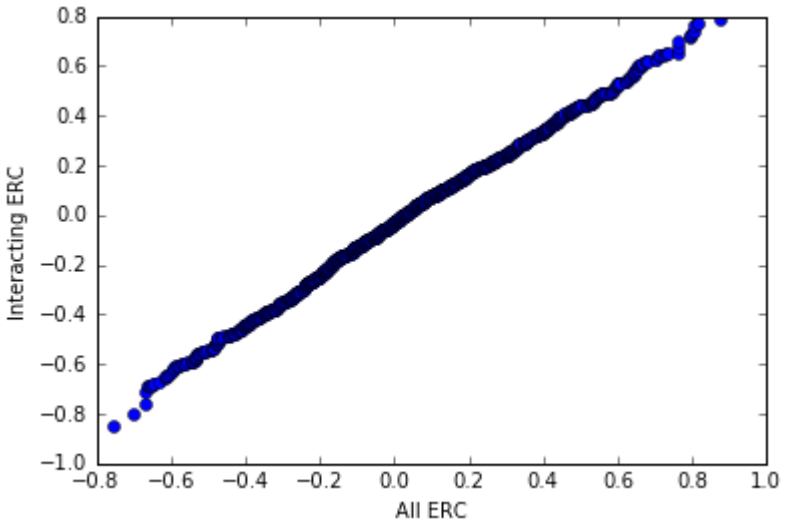
'erc std dev: 0.4526',

'erc tstat: 3.3121',

'erc pval: 0.0129']}


ERC Between All Yeast Proteins.png
Mean: -0.002752
200
150
100
50
0
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8


ERC Between Interacting Yeast Proteins.png
Mean: 0.039773
200
150
100
50
0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0


Interacting ERC
All ERC
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0

Q-Q plot

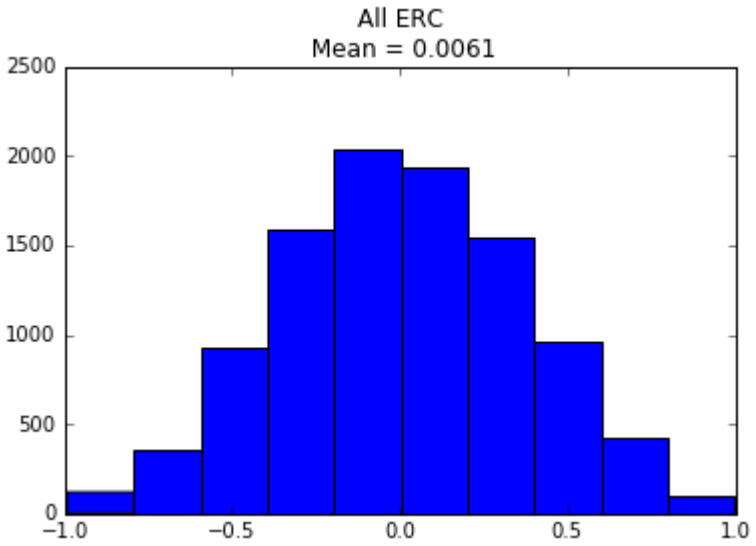
Biogrid

Tuesday, March 7, 2017 1:01 AM

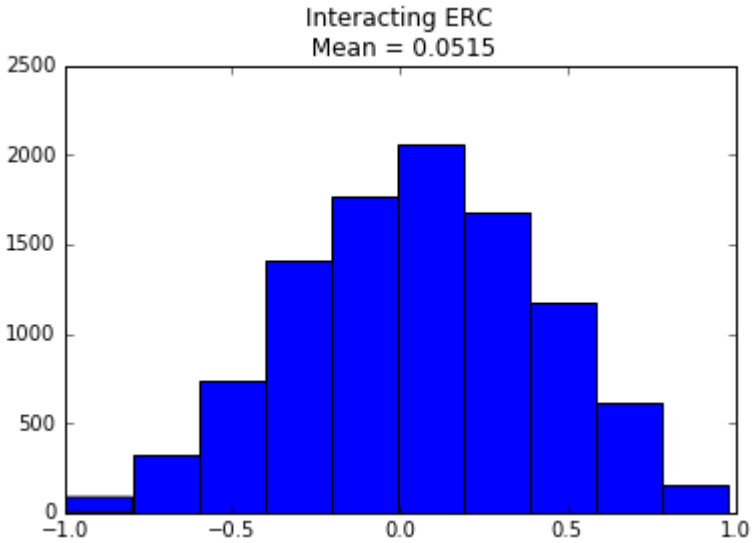
https://wiki.thebiogrid.org/doku.php/experimental_systems

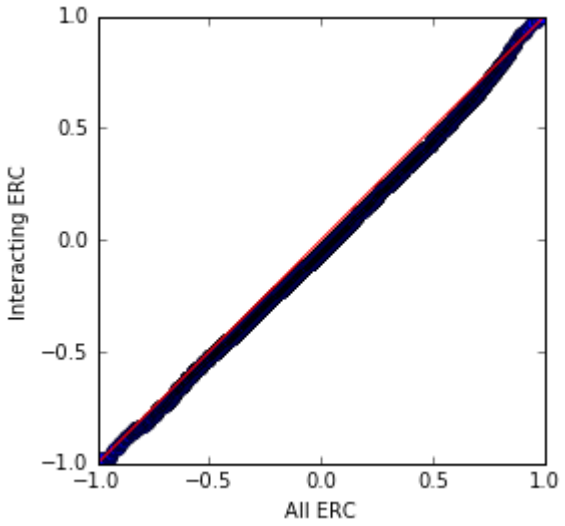
Biogrid ACMS results

Friday, March 24, 2017 4:45 AM

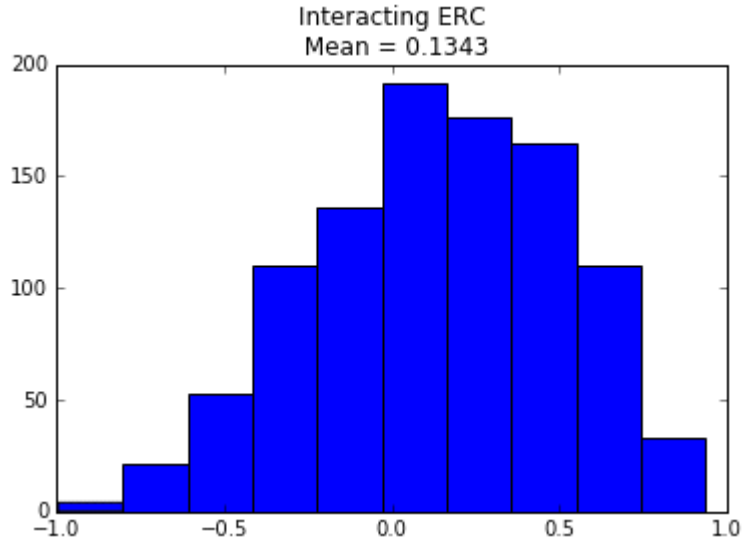


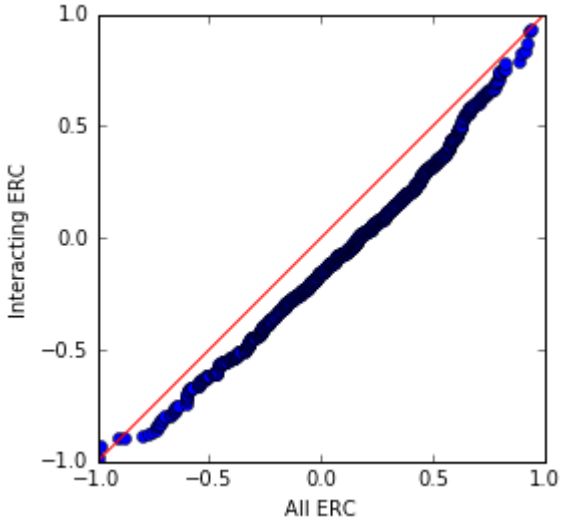
1 independent publication says interacting (32287 pairs):



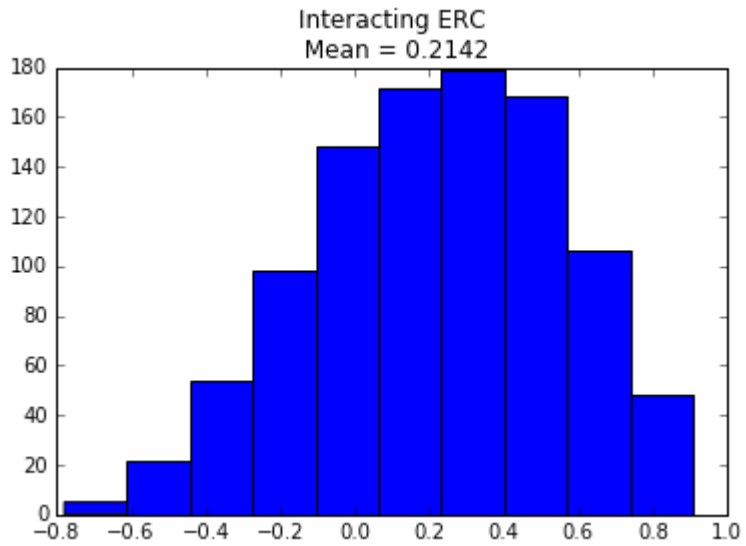


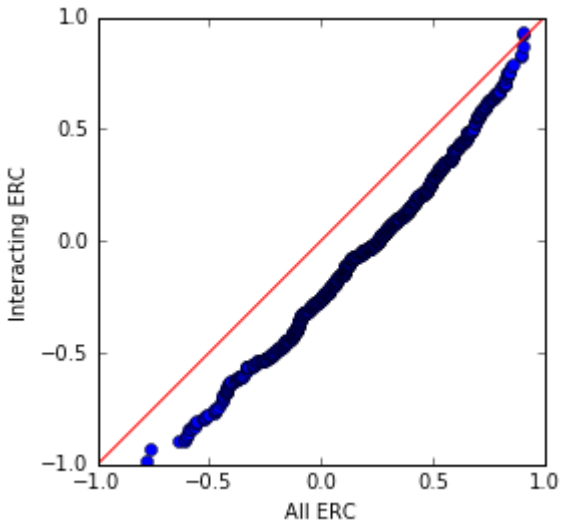
2 independent publications (6240):



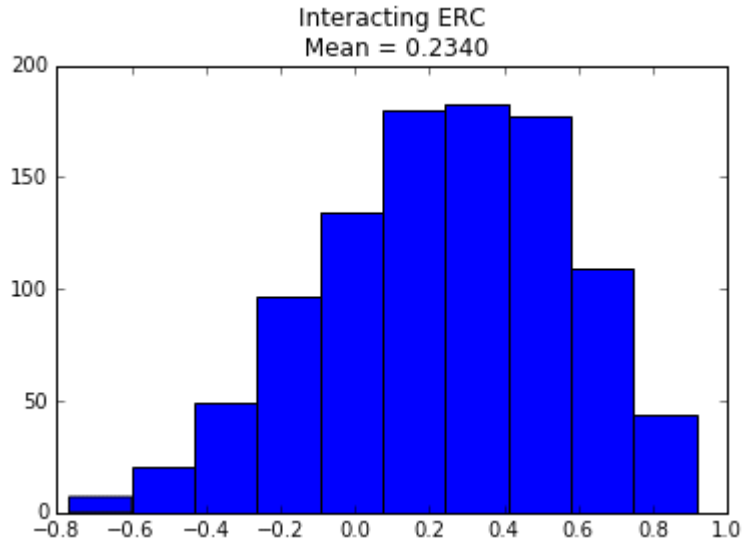


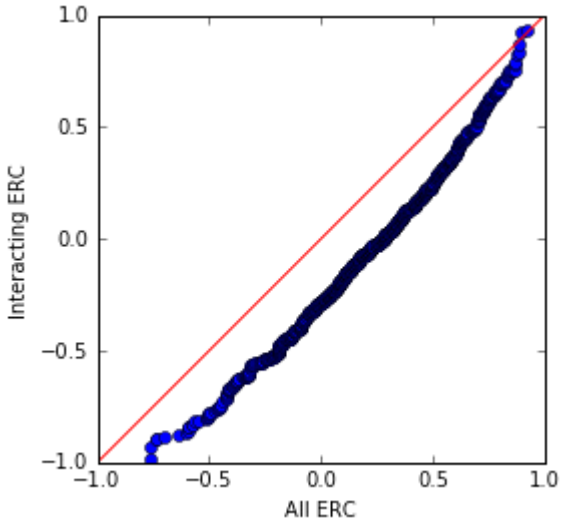
3 independent publications (2974):



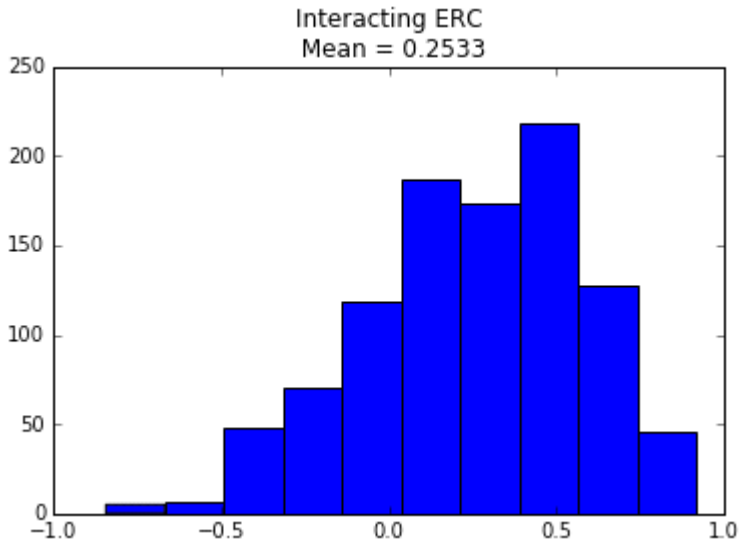


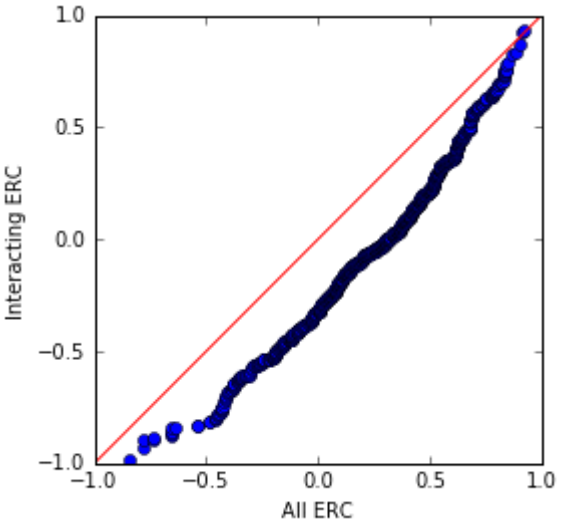
4 independent publications (1767):



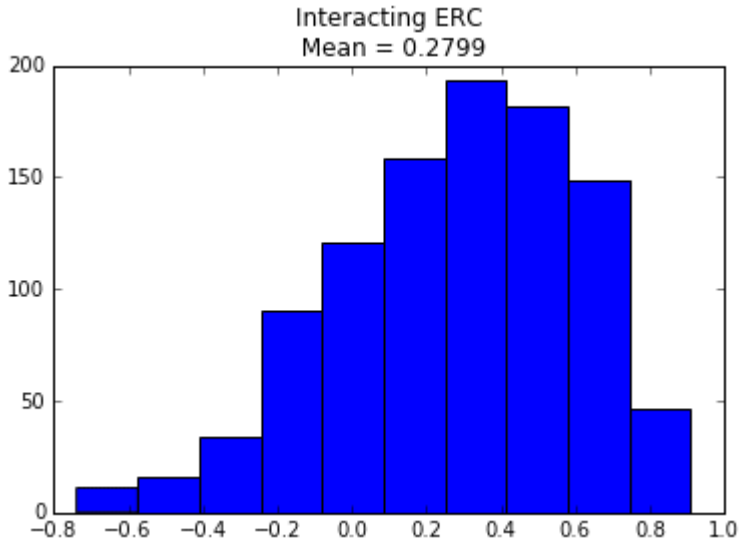


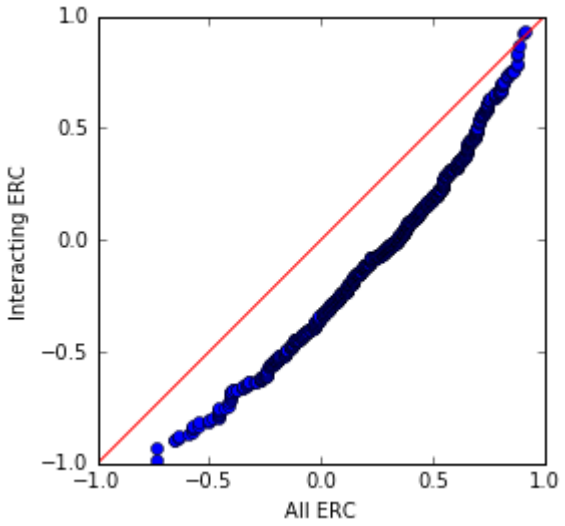
5 independent publications (1128):





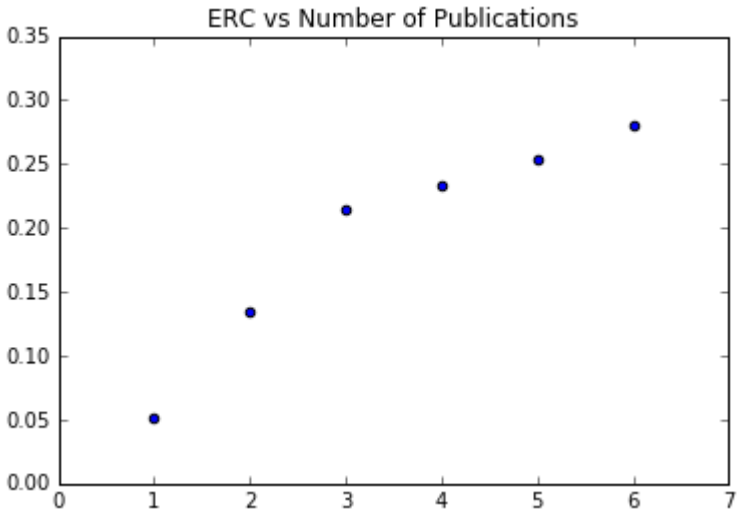
6 independent publications (701):




1.0
0.5
0.0
−0.5
−1.0
Interacting ERC
−1.0
−0.5
0.0
0.5
1.0
All ERC

Interaction Database, by number of publications

Monday, March 27, 2017 1:49 PM



Bonferroni

Monday, March 27, 2017 2:41 PM

Learn bonferroni
Distribution of P values

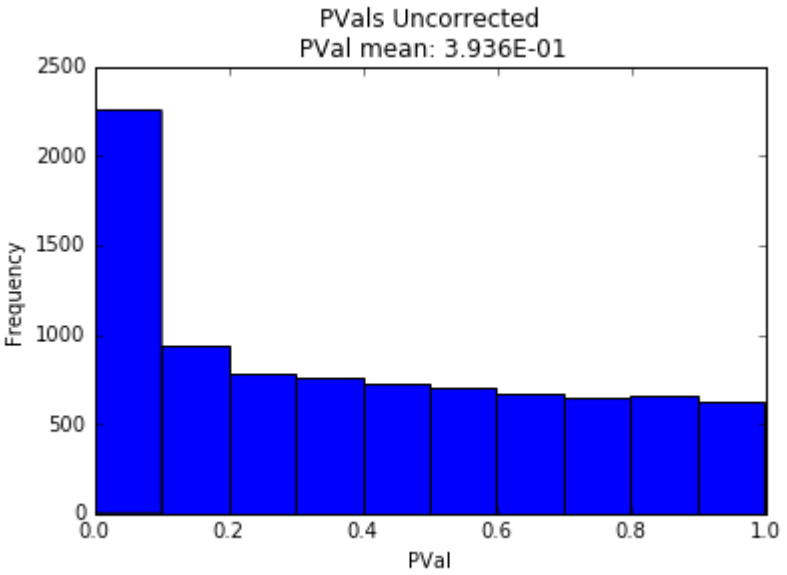
Q-Q plot uniform pvalue (if null model is true) probability.

Poster fair abstract

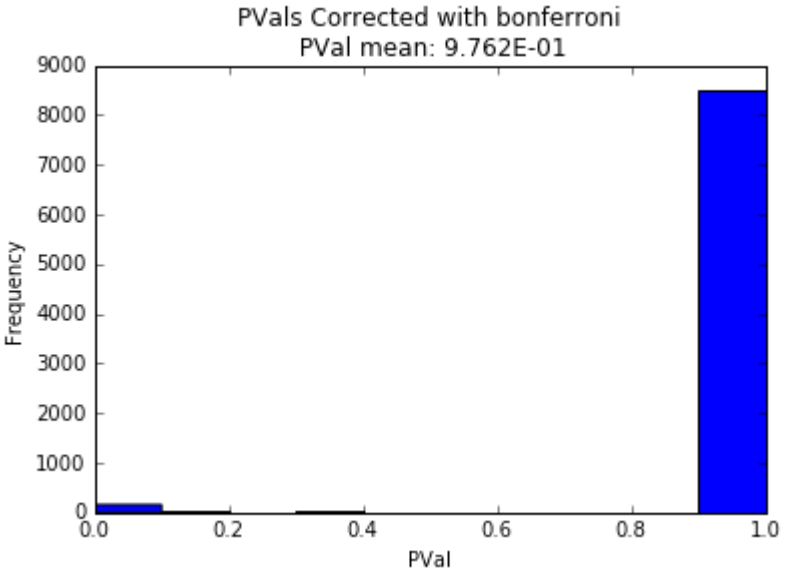
Bonferroni Corrected Pval Distribution

Monday, April 3, 2017 11:46 AM

1632 Values > 0.05



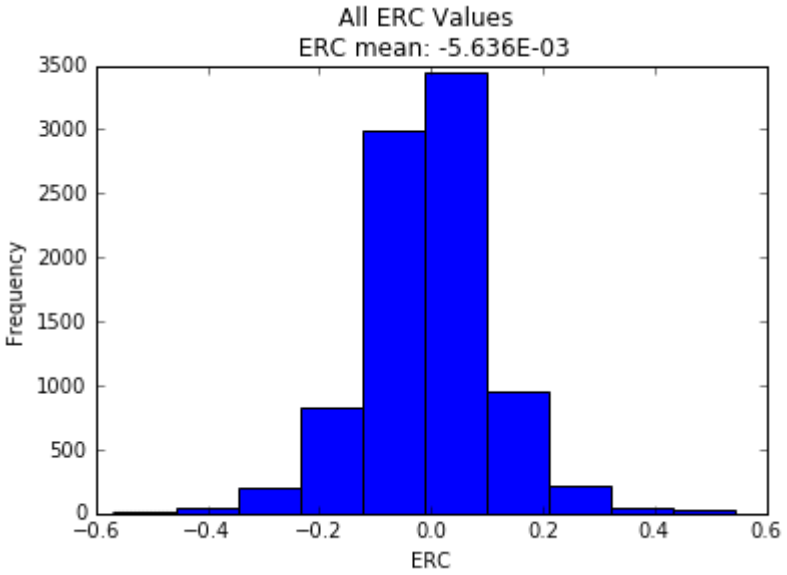
139 Values > 0.05



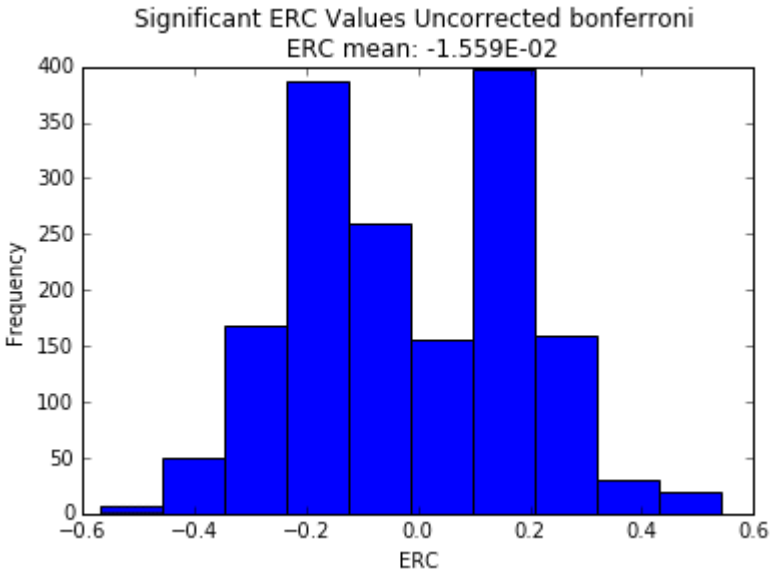
Significant ERC Distributions

Monday, April 3, 2017 12:12 PM

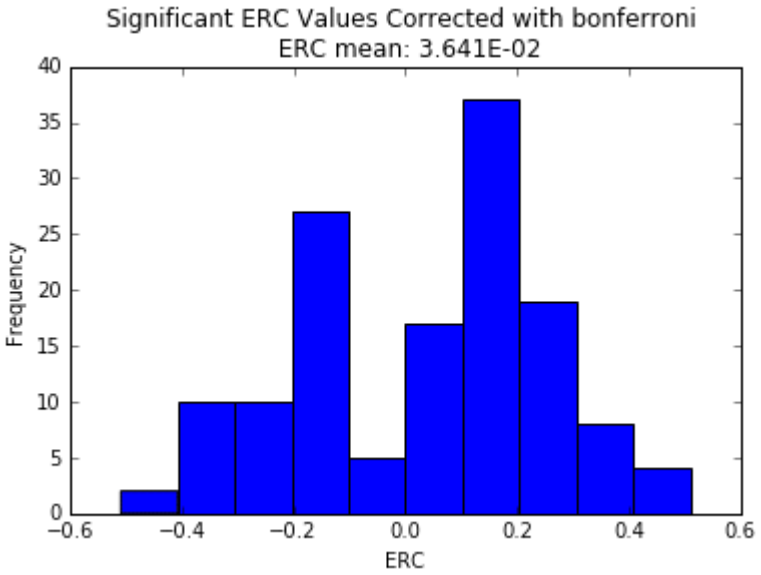
8778 Values:



1632 Values:



84 Values:



84 with > 0 ERC

68 > 0.1 ERC

32 > 0.2 ERC

11 > 0.3 ERC

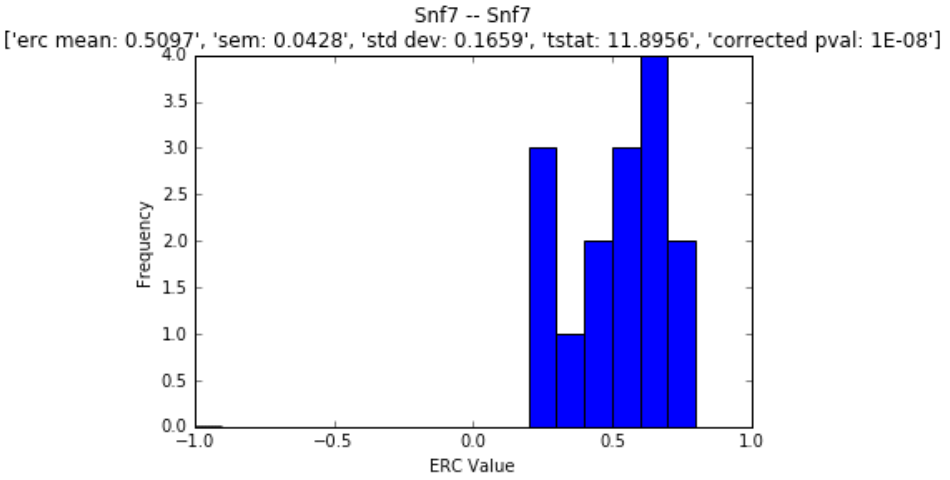
4 > 0.4 ERC

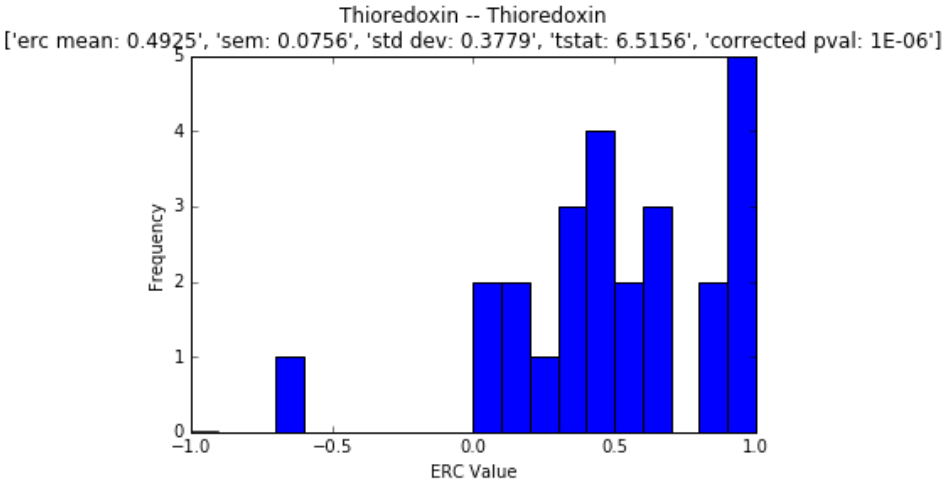
1 > 0.5 ERC

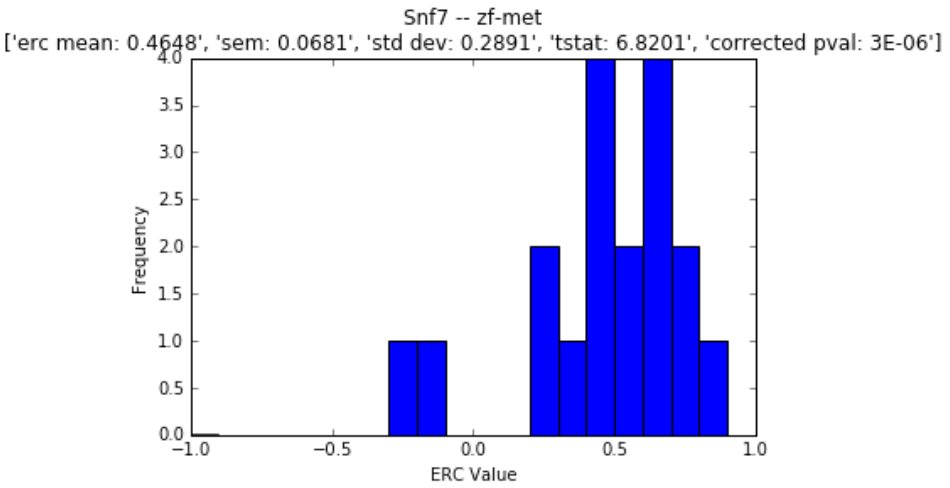
Top 10 Significant PVals ERC Values after Bonferroni

Monday, April 3, 2017 11:57 AM

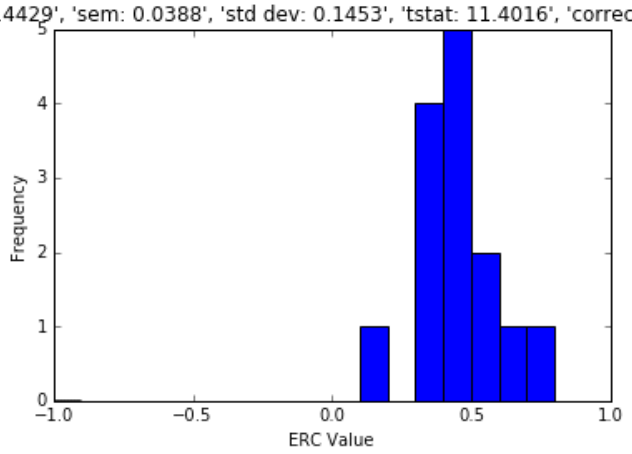
There were



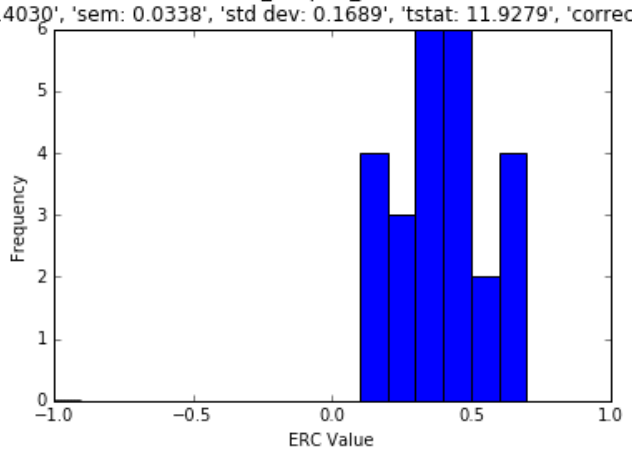




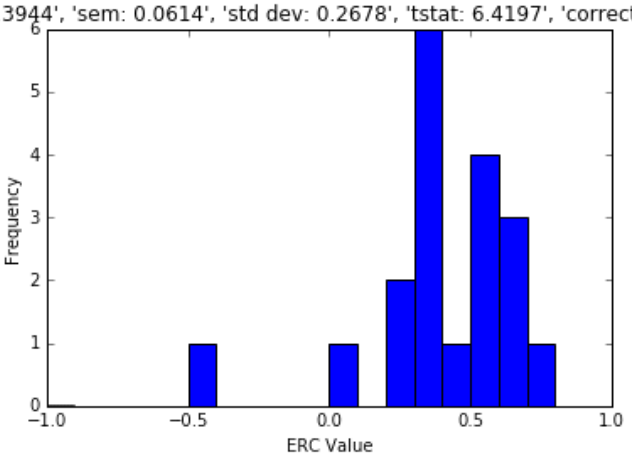
Brix -- Brix
['erc mean: 0.4429', 'sem: 0.0388', 'std dev: 0.1453', 'tstat: 11.4016', 'corrected pval: 4E-08']



Clat_adaptor_s -- Snf7
['erc mean: 0.4030', 'sem: 0.0338', 'std dev: 0.1689', 'tstat: 11.9279', 'corrected pval: 1E-11']

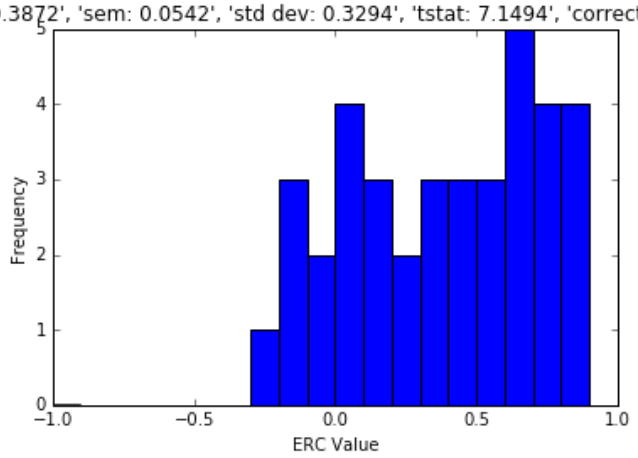


Histone -- Snf7
['erc mean: 0.3944', 'sem: 0.0614', 'std dev: 0.2678', 'tstat: 6.4197', 'corrected pval: 5E-06']



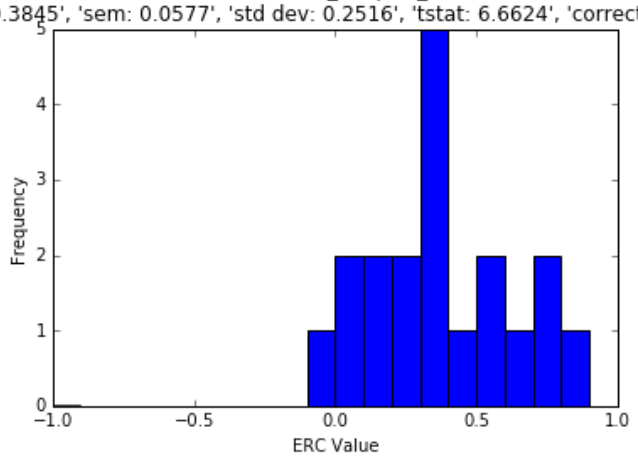
PX -- Snf7

['erc mean: 0.3872', 'sem: 0.0542', 'std dev: 0.3294', 'tstat: 7.1494', 'corrected pval: 2E-08']



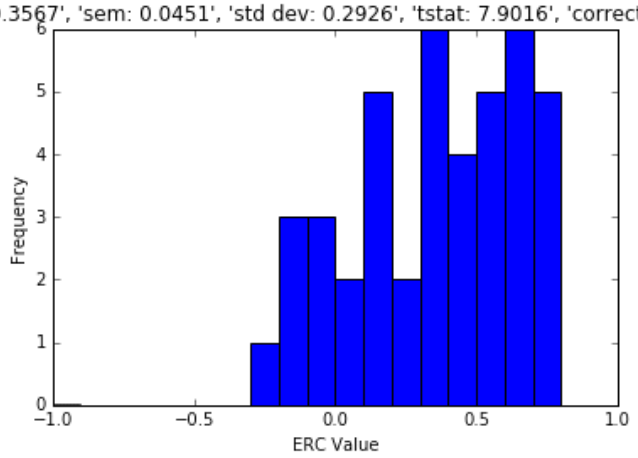
PX -- Clat_adaptor_s

['erc mean: 0.3845', 'sem: 0.0577', 'std dev: 0.2516', 'tstat: 6.6624', 'corrected pval: 3E-06']

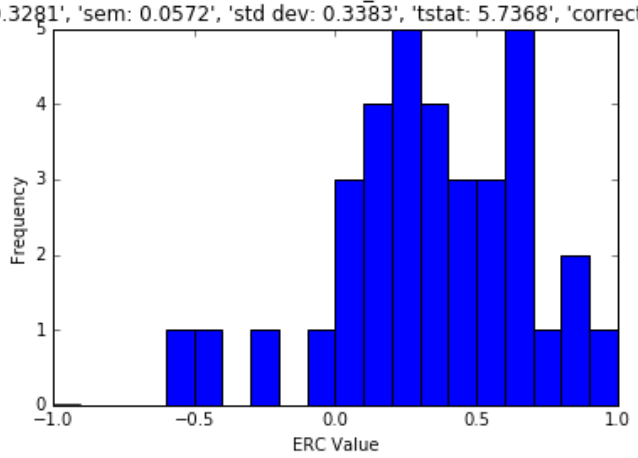


Snf7 -- Ras

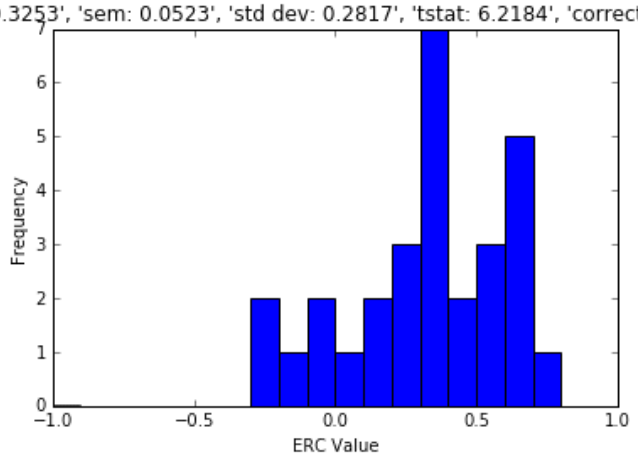
['erc mean: 0.3567', 'sem: 0.0451', 'std dev: 0.2926', 'tstat: 7.9016', 'corrected pval: 9E-10']



SUR7 -- adh_short
['erc mean: 0.3281', 'sem: 0.0572', 'std dev: 0.3383', 'tstat: 5.7368', 'corrected pval: 2E-06']

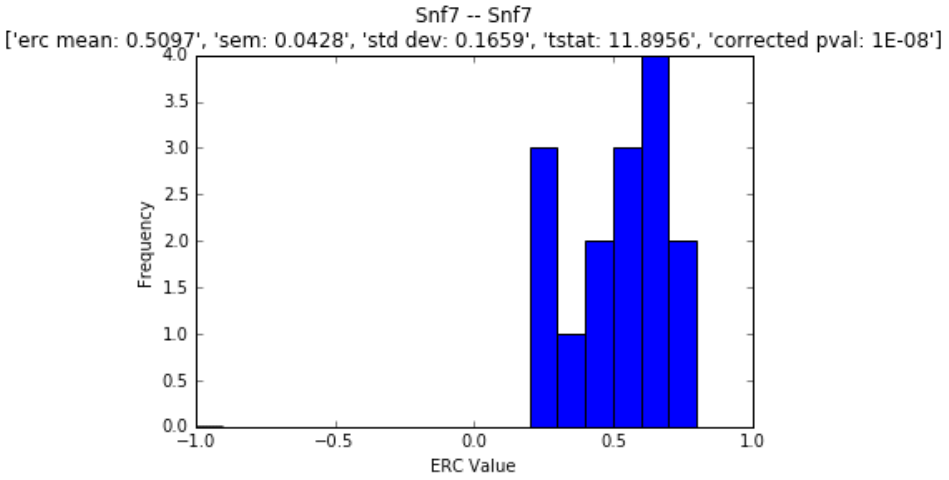


Snf7 -- PCI
['erc mean: 0.3253', 'sem: 0.0523', 'std dev: 0.2817', 'tstat: 6.2184', 'corrected pval: 1E-06']



Interpretation

Monday, April 3, 2017 12:24 PM

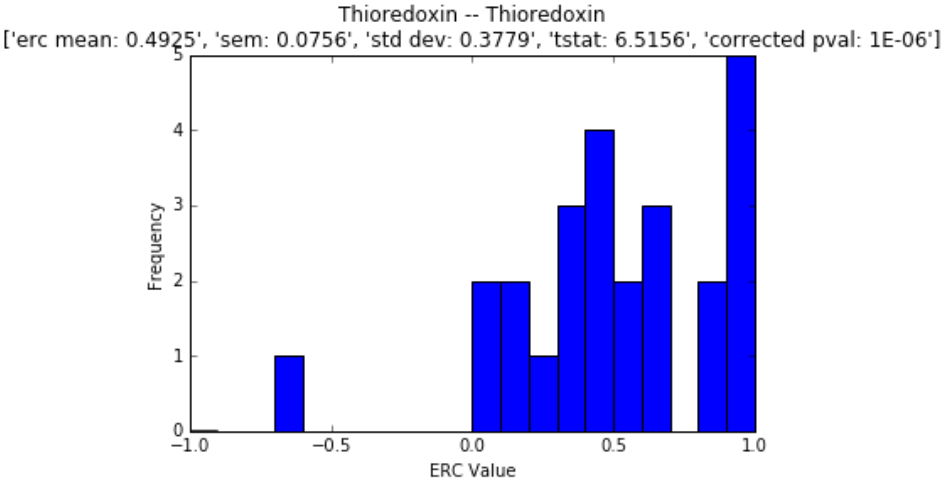


From <http://www.ebi.ac.uk/interpro/entry/IPR005024?q=snf7>

Snf7 family members are small coil-coiled proteins that share protein sequence similarity with budding yeast Snf7, which is part of the ESCRT-III complex that is required for endosome-mediated trafficking via multivesicular body (MVB) formation and sorting [PMID: 15086794].

Proteins in this entry also includes human CHMPs (charged multivesicular body proteins), budding yeast Did4/Did2 and Arabidopsis vacuolar protein sorting-associated proteins.

> We have found interactions between proteins within the ESCRT complexes, as expected from previous studies (**10,15,16**). Vps28p interacted with Vps37p (ESCRTI), Vps32p with Vps20p and Vps24p (ESCRTIII), and Vps25p, Vps22p, and Vps36p interacted with one another (ESCRTII). Vps32p interacted with itself, as did Vps46p/Did2p, suggesting that these proteins may form multimers. However, the oligomeric state of Vps32p remains unclear since a previous study found that His6-tagged Vps32p did not pull down endogenous Vps32p, suggesting that it is monomeric in solution (**16**). Vps27p also bound to Vps23p and Hse1p, as previously demonstrated (**18–20**). In addition we saw novel interactions of Hse1p with Vps23p and Rsp5p.
>
> From <http://onlinelibrary.wiley.com/doi/10.1111/j.1600-0854.2004.00169.x/full#t3n1>

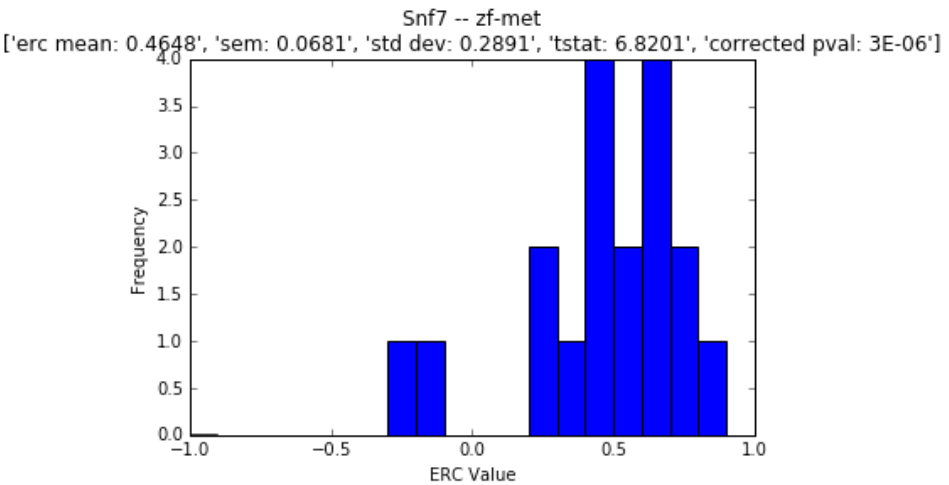


Thioredoxin domain (IPR013766)

From <http://www.ebi.ac.uk/interpro/entry/IPR013766?q=Thioredoxin>

Thioredoxins [PMID: 3896121, PMID: 2668278, PMID: 7788289, PMID: 7788290] are small disulphide-containing redox proteins that have been found in all the kingdoms of living organisms. Thioredoxin serves as a general protein disulphide oxidoreductase. It interacts with a broad range of proteins by a redox mechanism based on reversible oxidation of two cysteine thiol groups to a disulphide, accompanied by the transfer of two electrons and two protons. The net result is the covalent interconversion of a disulphide and a dithiol. In the NADPH-dependent protein disulphide reduction, thioredoxin reductase (TR) catalyses the reduction of oxidised thioredoxin (trx) by NADPH using FAD and its redox-active disulphide; reduced thioredoxin then directly reduces the disulphide in the substrate protein

http://onlinelibrary.wiley.com/doi/10.1110/ps.8.2.426/pdf

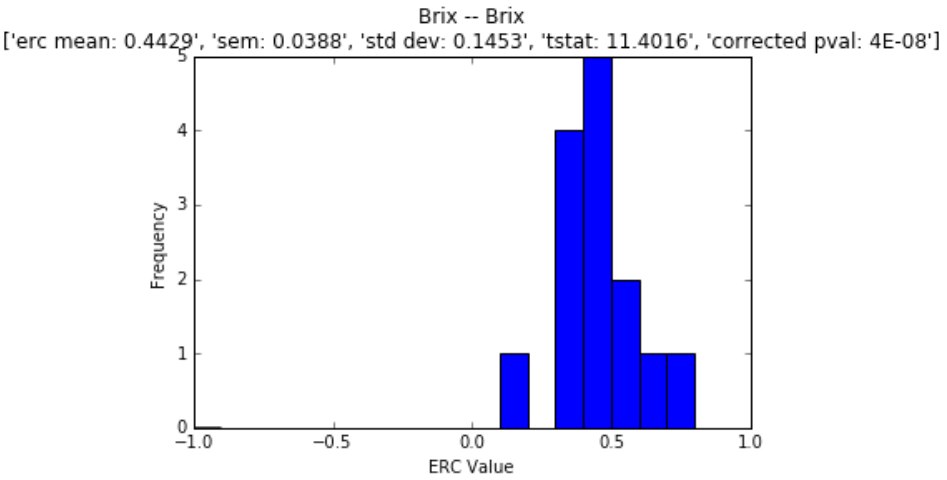


Zinc-finger of C2H2 type

From <http://pfam.xfam.org/family/zf-met>

This is a zinc-finger domain with the CxxCx(12)Hx(6)H motif, found in multiple copies in a wide range of proteins from plants to metazoans. Some member proteins, particularly those from plants, are annotated as being RNA-binding.

From <http://pfam.xfam.org/family/zf-met>



Brix domain (IPR007109)

Short name: Brix

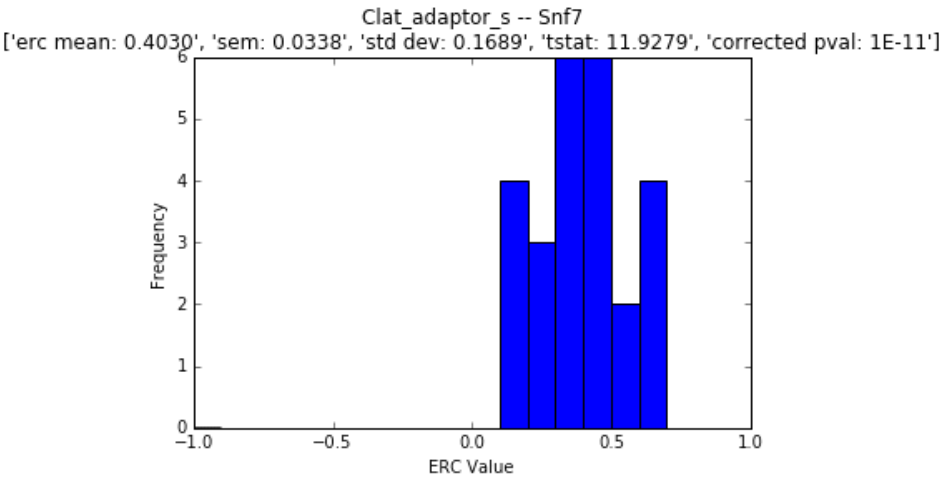
Analysis of the Brix (biogenesis of ribosomes in Xenopus) protein leaded to the identification of a region of 150-180 residues length, called the Brix domain, which is found in six protein families: one archaean family (I) including hypothetical proteins (one per genome); and five eukaryote families, each named according to a representative member and including close homologues of this prototype: (II) Peter Pan (D. melanogaster) and SSF1/2 (S.cerevisiae); (III) RPF1 (S. cerevisiae); (IV) IMP4 (S. cerevisiae); (V) Brix (X.laevis) and BRX1 (S. cerevisiae); and (VI) RPF2 (S.cerevisiae).

From <http://www.ebi.ac.uk/interpro/entry/IPR007109?q=brix>

Ribosome biogenesis protein BRX1 (IPR026532)

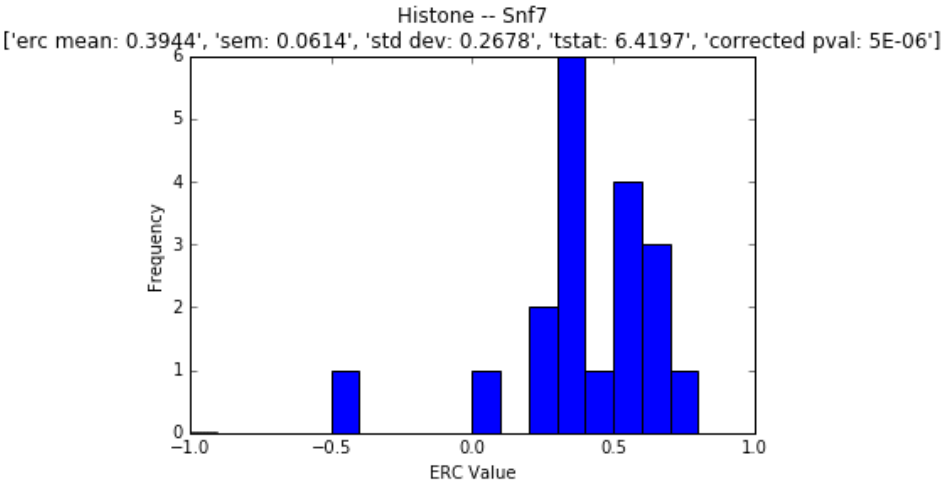
This family consists of BRX1 and homologues. In yeast, BRX1 is part of a complex that also includes RPF1, RPF2 and SSF1 or SSF2. It is required for biogenesis of the 60S ribosomal subunit [PMID: 11843177].

From <http://www.ebi.ac.uk/interpro/entry/IPR026532?q=brx1>



Clathrin Adaptor
endocytosis

https://www.ncbi.nlm.nih.gov/pubmed/11879634



Histone Domain?

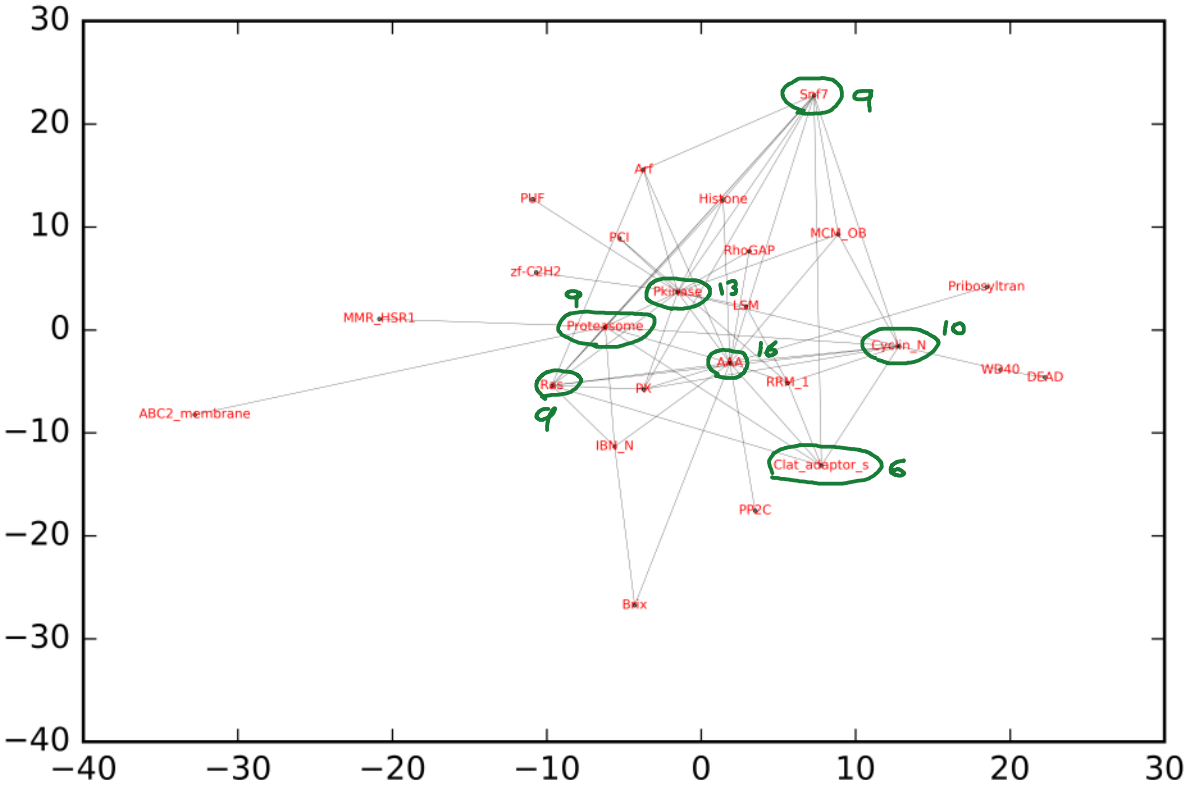
Notes

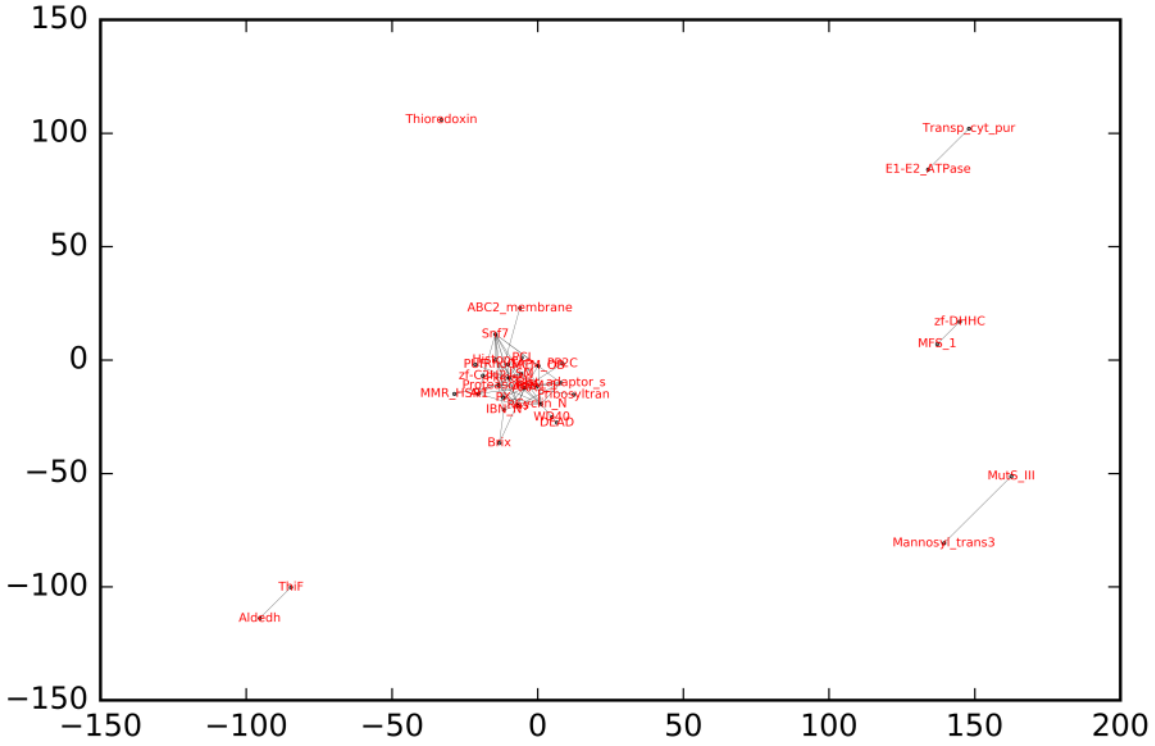
Monday, April 3, 2017 1:08 PM

Same protein remove values
Benjamini/Hochberg
One tailed test
Test against zero mean
Pfam references
Start writing
Network of inverse ERC pair distances, graphviz, networkx, degrees of relatedness, etc.
Network of protein ERCs

Domain Family ERC network

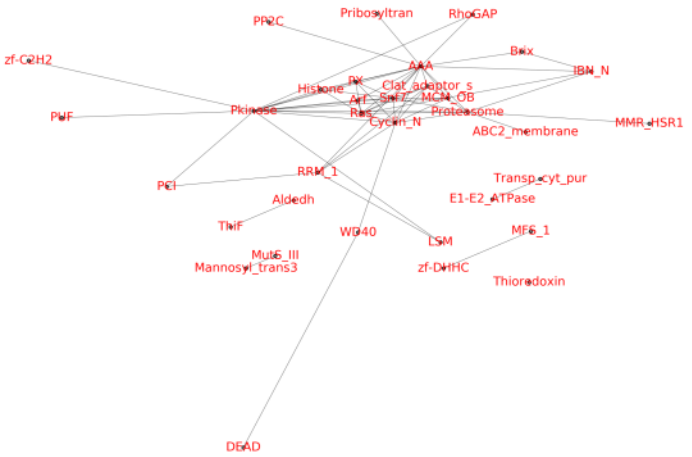
Monday, April 10, 2017 1:24 AM

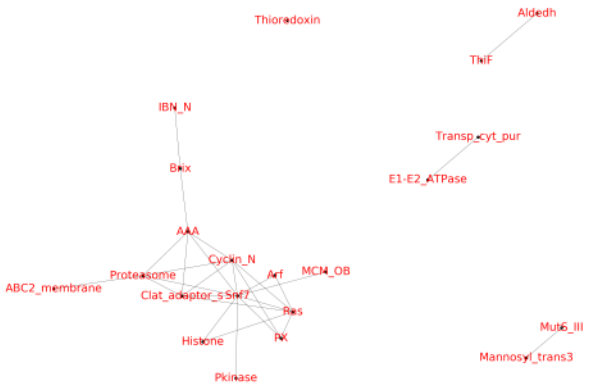



150
100
50
0
−50
−100
−150
−150
−100
−50
0
50
100
150
200
Thioredoxin
Transp_cyt_pur
E1-E2_ATPase
ABC2_membrane
Sof7
zf-DHHC
MFS_1
Mut6_III
Mannosyl_trans3
TMF
Aldedh

Inverse Network

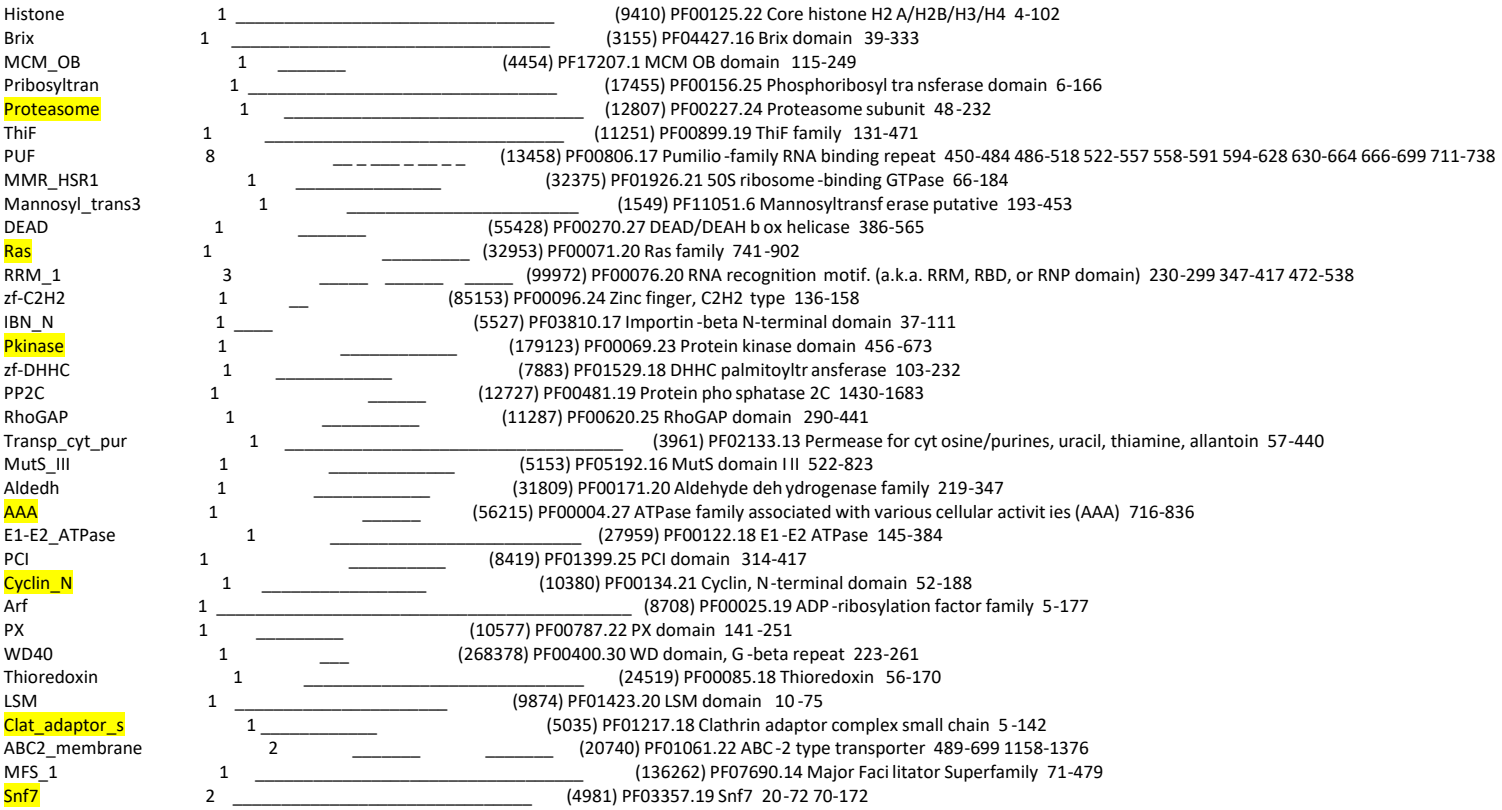
Monday, April 10, 2017 1:45 AM





Pfam descriptions

Monday, April 10, 2017 2:11 AM



Proteins

Monday, April 10, 2017 12:37 PM

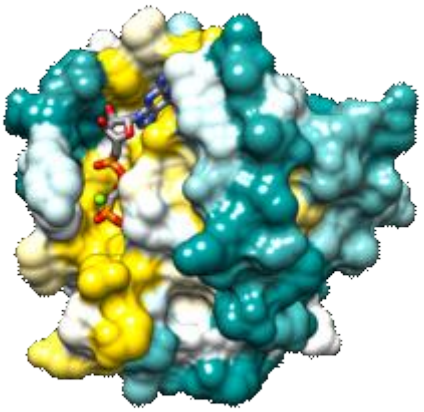
Proteasome subunit Provide feedback

The proteasome is a multisubunit structure that degrades proteins. Protein degradation is an essential component of regulation because proteins can become misfolded, damaged, or unnecessary. Proteasomes and their homologues vary greatly in complexity: from HslV (heat shock locus v), which is encoded by 1 gene in bacteria, to the eukaryotic 20S proteasome, which is encoded by more than 14 genes [1]. Recently evidence of two novel groups of bacterial proteasomes was proposed. The first is Anbu, which is sparsely distributed among cyanobacteria and proteobacteria [1]. The second is call beta-proteobacteria proteasome homologue (BPH) [1].

From <http://pfam.xfam.org/family/PF00227>

Ras subfamily Edit Wikipedia article

This article is about p21/Ras protein. For the p21/waf1 protein, see p21.



HRas structure PDB 121p, surface colored by conservation in Pfam seed alignment: gold, most conserved; dark cyan, least conserved.

Identifiers	
Symbol	Ras
Pfam	PF00071
InterPro	IPR020849
PROSITE	PDOC00017
SCOP	5p21
SUPERFAMILY	5p21
CDD	cd04138
Available protein structures:[show]	

Ras is a family of related proteins which is expressed in all animal cell lineages and organs. All Ras

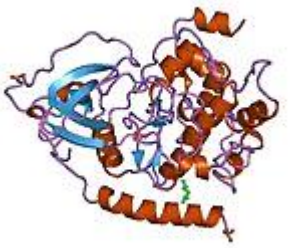
protein family members belong to a class of protein called small GTPase, and are involved in transmitting signals within cells (cellular signal transduction). Ras is the prototypical member of the Ras superfamily of proteins, which are all related in 3D structure and regulate diverse cell behaviours. When Ras is 'switched on' by incoming signals, it subsequently switches on other proteins, which ultimately turn on genes involved in cell growth, differentiation and survival. Mutations in ras genes can lead to the production of permanently activated Ras proteins. As a result, this can cause unintended and overactive signaling inside the cell, even in the absence of incoming signals.
Because these signals result in cell growth and division, overactive Ras signaling can ultimately lead to cancer.[1] The 3 Ras genes in humans (HRas, KRas, and NRas) are the most common oncogenes in human cancer; mutations that permanently activate Ras are found in 20% to 25% of all human tumors and up to 90% in certain types of cancer (e.g., pancreatic cancer).[2] For this reason, Ras inhibitors are being studied as a treatment for cancer and other diseases with Ras overexpression.

From <http://pfam.xfam.org/family/PF00071>

Includes sub-families Ras, Rab, Rac, Ral, Ran, Rap Ypt1 and more. Shares P-loop motif with GTP_EFTU, arf and myosin_head. See PF00009 PF00025 PF00063. As regards Rab GTPases, these are important regulators of vesicle formation, motility and fusion. They share a fold in common with all Ras GTPases: this is a six-stranded beta-sheet surrounded by five alpha-helices [1].

From <http://pfam.xfam.org/family/PF00071>

Protein kinase domain Edit Wikipedia article

Protein kinase domain	
 Structure of the catalytic subunit of cAMP-dependent protein kinase.[1]	
Identifiers	
Symbol	Pkinase
Pfam	PF00069
InterPro	IPR000719
SMART	TyrKc
PROSITE	PDOC00100
SCOP	1apm
SUPERFAMILY	1apm
OPM superfamily	417
OPM protein	2w5a

CDD	cd00180
Available protein structures:[show]	

The **protein kinase domain** is a structurally conserved protein domain containing the catalytic function of protein kinases.[2][3][4] Protein kinases are a group of enzymes that move a phosphate group onto proteins, in a process called phosphorylation. This functions as an on/off switch for many cellular processes, including metabolism, transcription, cell cycle progression, cytoskeletal rearrangement and cell movement, apoptosis, and differentiation. They also function in embryonic development, physiological responses, and in the nervous and immune system. Abnormal phosphorylation causes many human diseases, including cancer, and drugs that affect phosphorylation can treat those diseases.[5]
Protein kinases possess a catalytic subunit which transfers the gamma phosphate from nucleoside triphosphates (often ATP) to one or more amino acid residues in a protein substrate side chain, resulting in a conformational change affecting protein function. These enzymes fall into two broad classes, characterised with respect to substrate specificity: serine/threonine specific and tyrosine specific.[6]

From <http://pfam.xfam.org/family/PF00069>

AAA proteins Edit Wikipedia article
For other uses, see AAA (disambiguation).

ATPases associated with diverse cellular activities (AAA)	
 Structure of N-ethylmaleimide-sensitive factor.[1]	
Identifiers	
Symbol	AAA
Pfam	PF00004
Pfam clan	CL0023
InterPro	IPR003959
PROSITE	PDOC00572
SCOP	1nsf
SUPERFAMILY	1nsf
CDD	cd00009
Available protein structures:[show]	

AAA or **AAA+** is an abbreviation for **ATPases Associated with diverse cellular Activities**. They share a common conserved module of approximately 230 amino acid residues. This is a large, functionally diverse protein family belonging to the AAA superfamily of ring-shaped P-loop NTPases, which exert their activity through the energy-dependent remodeling or translocation of macromolecules.[2][3]
AAA proteins couple chemical energy provided by ATP hydrolysis to conformational changes which are

transduced into mechanical force exerted on a macromolecular substrate.[4]
AAA proteins are functionally and organizationally diverse, and vary in activity, stability, and mechanism.[4] Members of the AAA family are found in all organisms[5] and they are essential for many cellular functions. They are involved in processes such as DNA replication, protein degradation, membrane fusion, microtubule severing, peroxisome biogenesis, signal transduction and the regulation of gene expression.

From <http://pfam.xfam.org/family/PF00004>

InterPro entry IPR003959

AAA ATPases (ATPases Associated with diverse cellular Activities) form a large protein family and play a number of roles in the cell including cell-cycle regulation, protein proteolysis and disaggregation, organelle biogenesis and intracellular transport. Some of them function as molecular chaperones, subunits of proteolytic complexes or independent proteases (FtsH, Lon). They also act as DNA helicases and transcription factors [PUBMED:17201069].
AAA ATPases belong to the AAA+ superfamily of ringshaped P-loop NTPases, which act via the energy-dependent unfolding of macromolecules [PUBMED:15037233, PUBMED:16828312]. There are six major clades of AAA domains (proteasome subunits, metalloproteases, domains D1 and D2 of ATPases with two AAA domains, the MSP1/katanin/spastin group and BCS1 and it homologues), as well as a number of deeply branching minor clades [PUBMED:15037233].
They assemble into oligomeric assemblies (often hexamers) that form a ring-shaped structure with a central pore. These proteins produce a molecular motor that couples ATP binding and hydrolysis to changes in conformational states that act upon a target substrate, either translocating or remodelling it [PUBMED:16919475].
They are found in all living organisms and share the common feature of the presence of a highly conserved AAA domain called the AAA module. This domain is responsible for ATP binding and hydrolysis. It contains 200-250 residues, among them there are two classical motifs, Walker A (GX4GKT) and Walker B (HyDE) [PUBMED:17201069].
The functional variety seen between AAA ATPases is in part due to their extensive number of accessory domains and factors, and to their variable organisation within oligomeric assemblies, in addition to changes in key functional residues within the ATPase domain itself.

From <http://pfam.xfam.org/family/PF00004>

InterPro entry IPR006671

Cyclins are eukaryotic proteins that play an active role in controlling nuclear cell division cycles [PUBMED:12910258], and regulate cyclin dependent kinases (CDKs). Cyclins, together with the p34 (cdc2) or cdk2 kinases, form the Maturation Promoting Factor (MPF). There are two main groups of cyclins, G1/S cyclins, which are essential for the control of the cell cycle at the G1/S (start) transition, and G2/M cyclins, which are essential for the control of the cell cycle at the G2/M (mitosis) transition. G2/M cyclins accumulate steadily during G2 and are abruptly destroyed as cells exit from mitosis (at the end of the M-phase). In most species, there are multiple forms of G1 and G2 cyclins. For example, in vertebrates, there are two G2 cyclins, A and B, and at least three G1 cyclins, C, D, and E.
Cyclin homologues have been found in various viruses, including Saimiriine herpesvirus 2 (Herpesvirus saimiri) and Human herpesvirus 8 (HHV-8) (Kaposi's sarcoma-associated herpesvirus). These viral homologues differ from their cellular counterparts in that the viral proteins have gained new functions and eliminated others to harness the cell and benefit the virus [PUBMED:11056549].
Cyclins contain two domains of similar all-alpha fold, of which this entry is associated with the N-terminal domain.

From <http://pfam.xfam.org/family/PF00134>

Clathrin adaptor complex small chain

From <http://pfam.xfam.org/family/PF01217.15>

InterPro entry IPR022775

Proteins synthesized on the ribosome and processed in the endoplasmic reticulum are transported from the Golgi apparatus to the trans-Golgi network (TGN), and from there via small carrier vesicles to their final destination compartment. This traffic is bidirectional, to ensure that proteins required to form vesicles are recycled. Vesicles have specific coat proteins (such as clathrin or coatomer) that are important for cargo selection and direction of transfer [PUBMED:15261670].

Clathrin coats contain both clathrin and adaptor complexes that link clathrin to receptors in coated vesicles. Clathrin-associated protein complexes are believed to interact with the cytoplasmic tails of membrane proteins, leading to their selection and concentration. The two major types of clathrin adaptor complexes are the heterotetrameric adaptor protein (AP) complexes, and the monomeric GGA (Golgi-localising, Gamma-adaptin ear domain homology, ARF-binding proteins) adaptors [PUBMED:17449236]. All AP complexes are heterotetramers composed of two large subunits (adaptins), a medium subunit (mu) and a small subunit (sigma). Each subunit has a specific function. Adaptin subunits recognise and bind to clathrin through their hinge region (clathrin box), and recruit accessory proteins that modulate AP function through their C-terminal appendage domains. By contrast, GGAs are monomers composed of four domains, which have functions similar to AP subunits: an N-terminal VHS (Vps27p/Hrs/Stam) domain, a GAT (GGA and Tom1) domain, a hinge region, and a C-terminal GAE (gamma-adaptin ear) domain. The GAE domain is similar to the AP gamma-adaptin ear domain, being responsible for the recruitment of accessory proteins that regulate clathrin-mediated endocytosis [PUBMED:12858162].

While clathrin mediates endocytic protein transport from ER to Golgi, coatomers (COPI, COPII) primarily mediate intra-Golgi transport, as well as the reverse Golgi to ER transport of dilysine-tagged proteins [PUBMED:14690497]. Coatomers reversibly associate with Golgi (non-clathrin-coated) vesicles to mediate protein transport and for budding from Golgi membranes [PUBMED:17041781]. Coatomer complexes are hetero-oligomers composed of at least an alpha, beta, beta', gamma, delta, epsilon and zeta subunits.

This entry represents the small sigma and mu subunits of various adaptins from different AP clathrin adaptor complexes (including AP1, AP2, AP3 and AP4), and the zeta and delta subunits of various coatomer (COP) adaptors. The small sigma subunit of AP proteins have been characterised in several species [PUBMED:8157009, PUBMED:8373805, PUBMED:8157009, PUBMED:9002613]. The sigma subunit plays a role in protein sorting in the late-Golgi/trans-Golgi network (TGN) and/or endosomes. The zeta subunit of coatomers (zeta-COP) is required for coatomer binding to Golgi membranes and for coat-vesicle assembly [PUBMED:8276893, PUBMED:14729954].

From <http://pfam.xfam.org/family/PF01217.15>

Snf7 Provide feedback

This family of proteins are involved in protein sorting and transport from the endosome to the vacuole/lysosome in eukaryotic cells. Vacuoles/lysosomes play an important role in the degradation of both lipids and cellular proteins. In order to perform this degradative function, vacuoles/lysosomes contain numerous hydrolases which have been transported in the form of inactive precursors via the biosynthetic pathway and are proteolytically activated upon delivery to the vacuole/lysosome. The delivery of transmembrane proteins, such as activated cell surface receptors to the lumen of the vacuole/lysosome, either for degradation/downregulation, or in the case of hydrolases, for proper

localisation, requires the formation of multivesicular bodies (MVBs). These late endosomal structures are formed by invaginating and budding of the limiting membrane into the lumen of the compartment. During this process, a subset of the endosomal membrane proteins is sorted into the forming vesicles. Mature MVBs fuse with the vacuole/lysosome, thereby releasing cargo containing vesicles into its hydrolytic lumen for degradation. Endosomal proteins that are not sorted into the intralumenal MVB vesicles are either recycled back to the plasma membrane or Golgi complex, or remain in the limiting membrane of the MVB and are thereby transported to the limiting membrane of the vacuole/lysosome as a consequence of fusion. Therefore, the MVB sorting pathway plays a critical role in the decision between recycling and degradation of membrane proteins [1]. A few archaeal sequences are also present within this family.

From <http://pfam.xfam.org/family/PF03357>

Interpretation

Monday, April 10, 2017 12:51 PM

Transport: SNF7, Clathrin,

Cell division: Cyclin,

Signaling/regulatory: ATPases Associated with diverse cellular Activities(gene regulation, protein degradation,membrane fusion), Protein kinase, Ras, (Proteasome ?)

Travis: highly conserved, housekeeping genes, consistent low evolutionary rate-->low erc

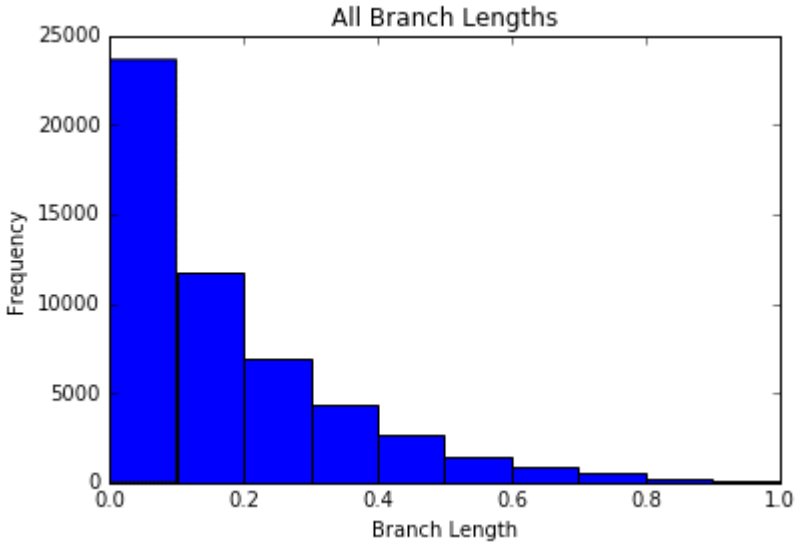
Really common, survive bonferroni test

ENR2

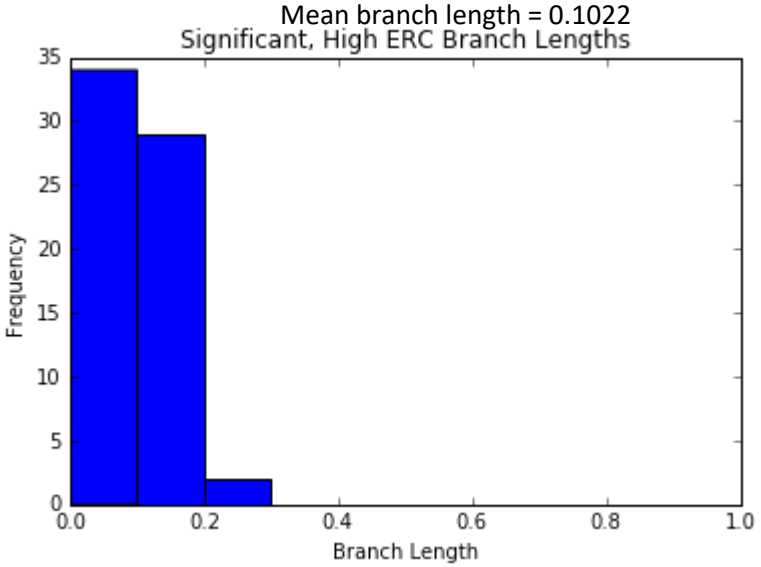
Compare Branch Lengths

Monday, April 10, 2017 7:19 PM

Mean branch length = 0.1955



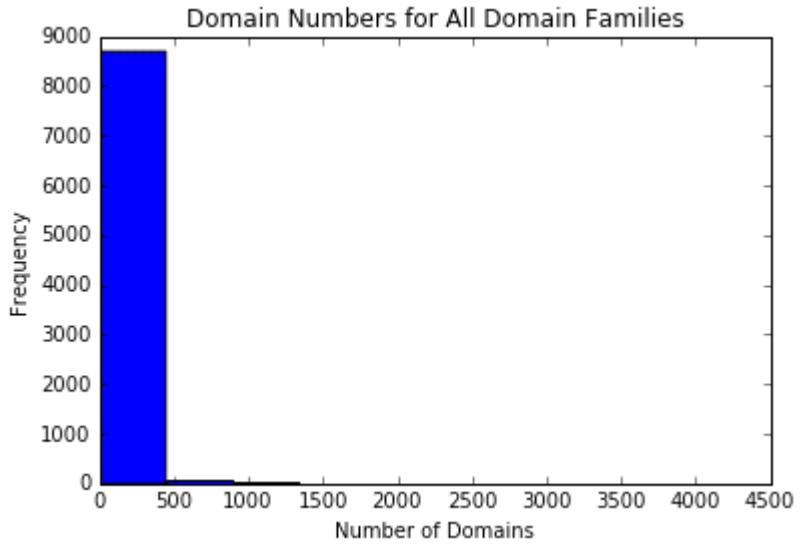
Mean branch length = 0.1022



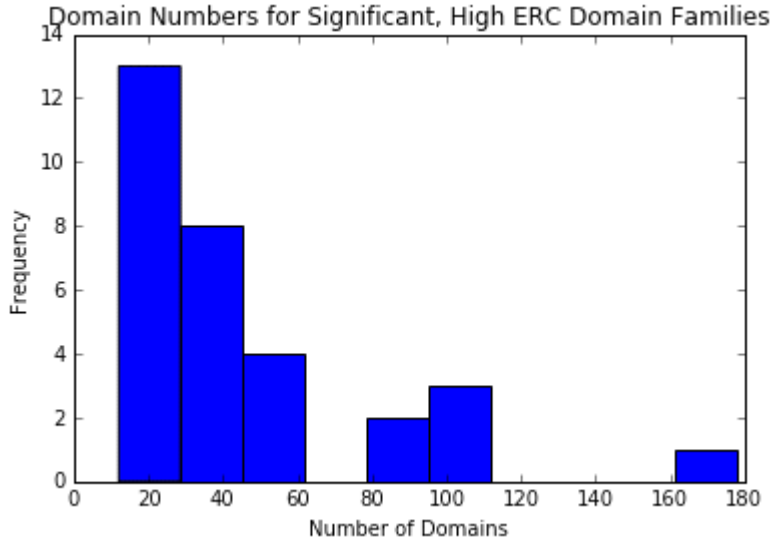
Compare Data Numbers

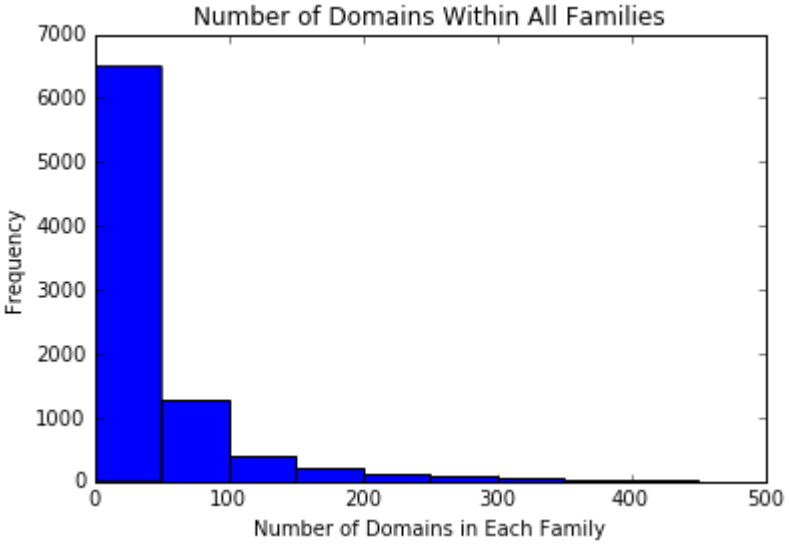
Monday, April 10, 2017 7:36 PM

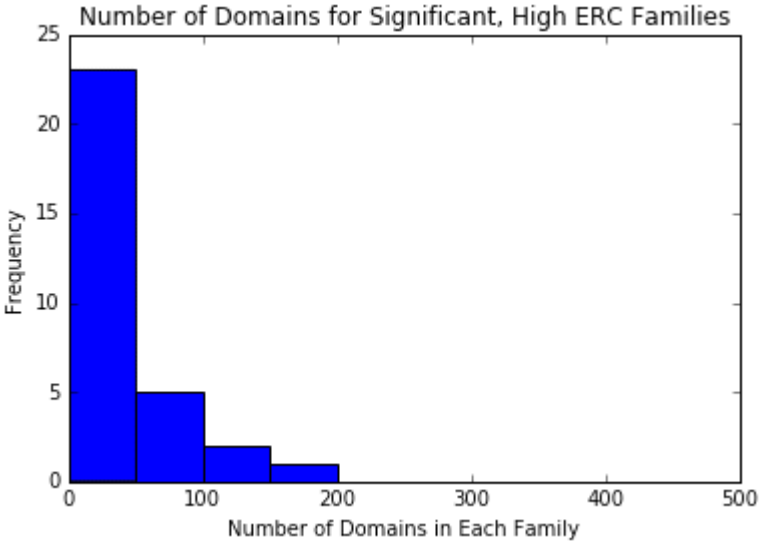
Mean Data Number: 55



Mean Sig Data Number: 45

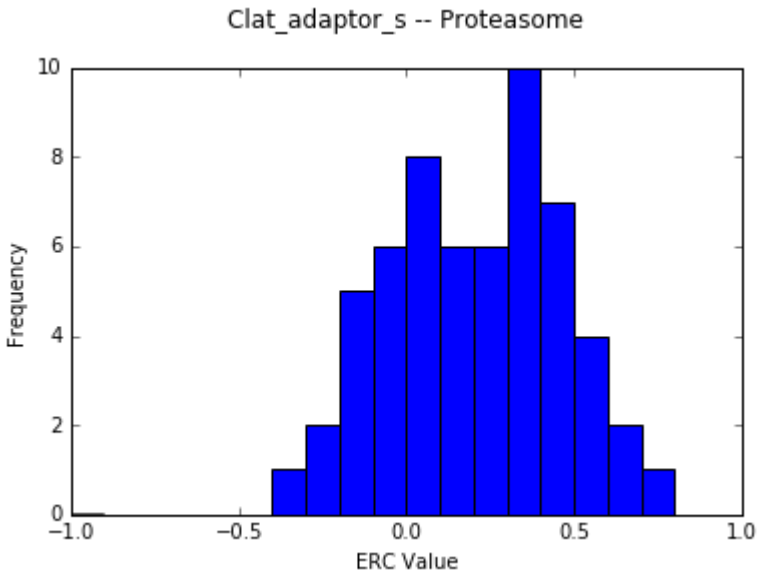


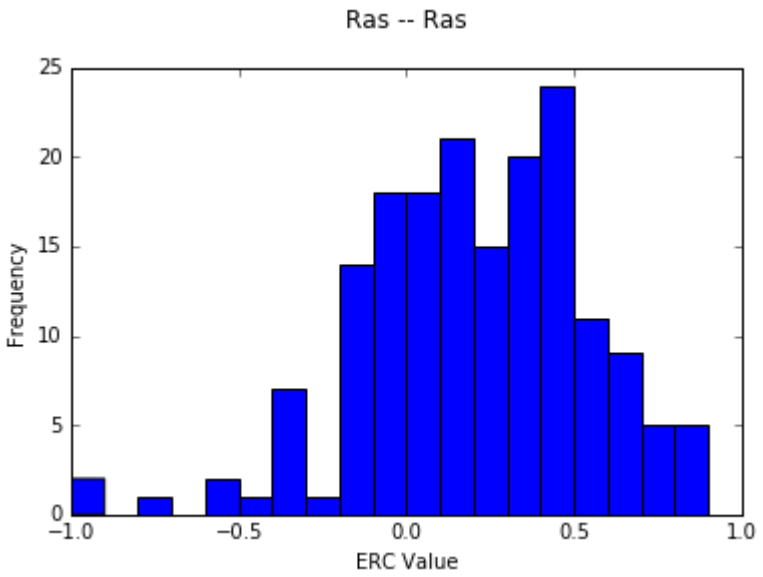

Number of Domains Within All Families
Frequency
Number of Domains in Each Family
0
1000
2000
3000
4000
5000
6000
7000
0
100
200
300
400
500

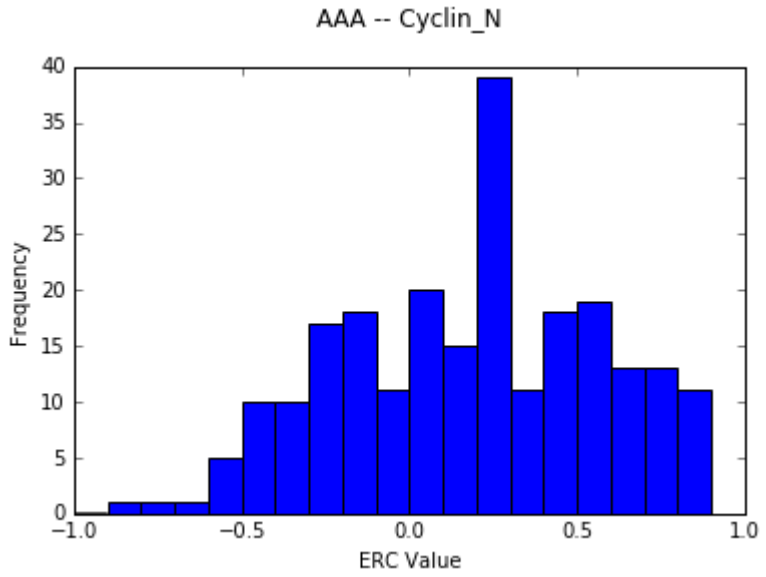

Number of Domains for Significant, High ERC Families
Frequency
Number of Domains in Each Family
0
5
10
15
20
25
0
100
200
300
400
500

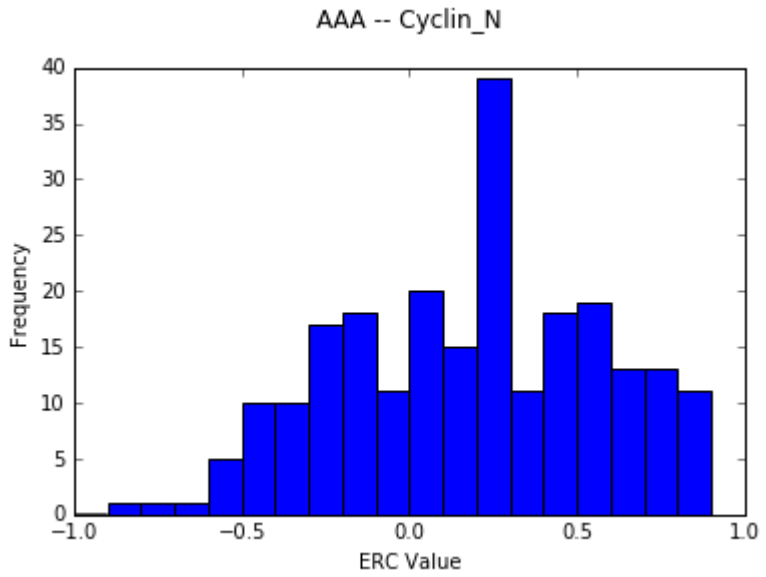
Draft figs

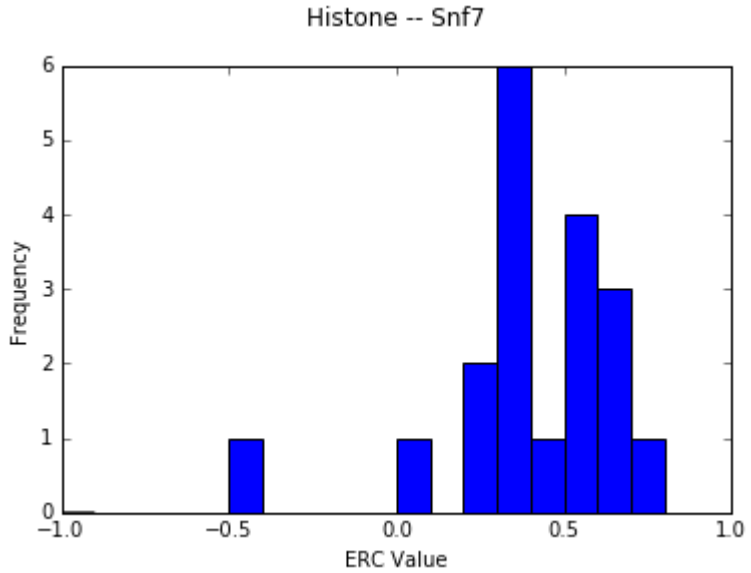
Monday, April 17, 2017 2:01 PM

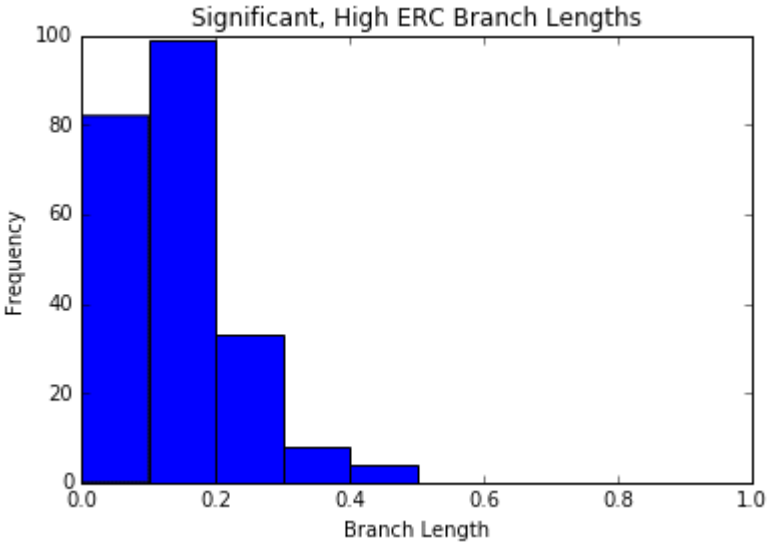





AAA -- Cyclin_N
Frequency
ERC Value
0
5
10
15
20
25
30
35
40
−1.0
−0.5
0.0
0.5
1.0


AAA -- Cyclin_N
Frequency
ERC Value
0
5
10
15
20
25
30
35
40
−1.0
−0.5
0.0
0.5
1.0


Histone -- Snf7
Frequency
ERC Value
0
1
2
3
4
5
6
−1.0
−0.5
0.0
0.5
1.0

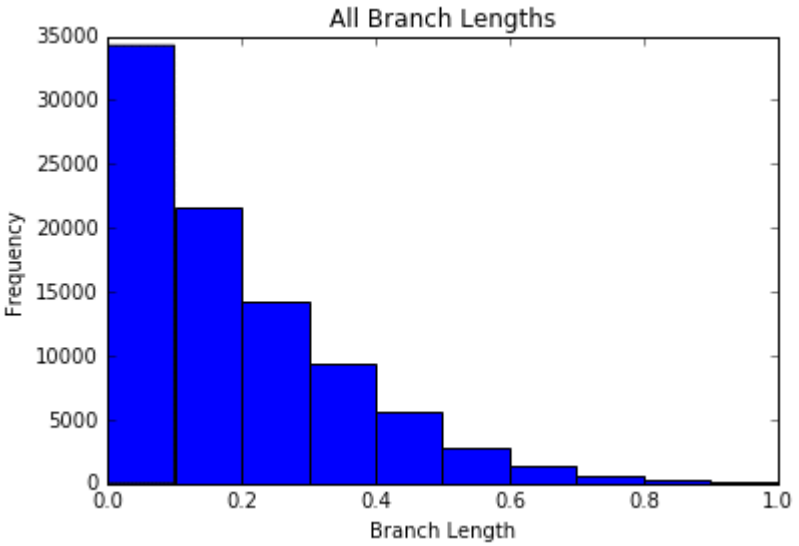

Significant, High ERC Branch Lengths
Frequency
Branch Length
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0

Mean = 0.1388

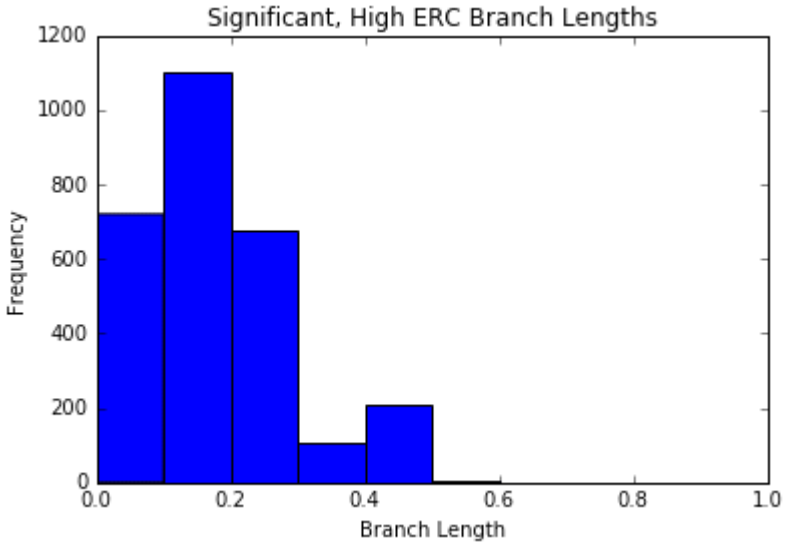
Whole Protein

Thursday, April 20, 2017 10:34 AM

Mean = 0.1918



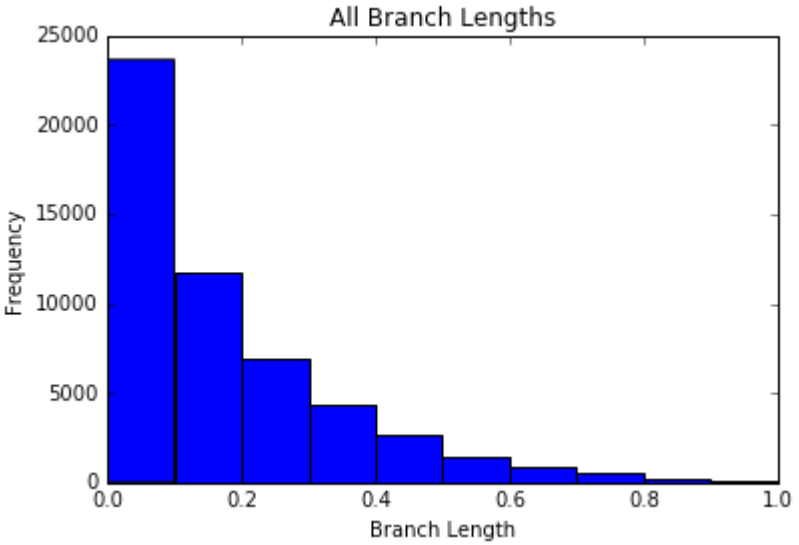
Mean = 0.1818



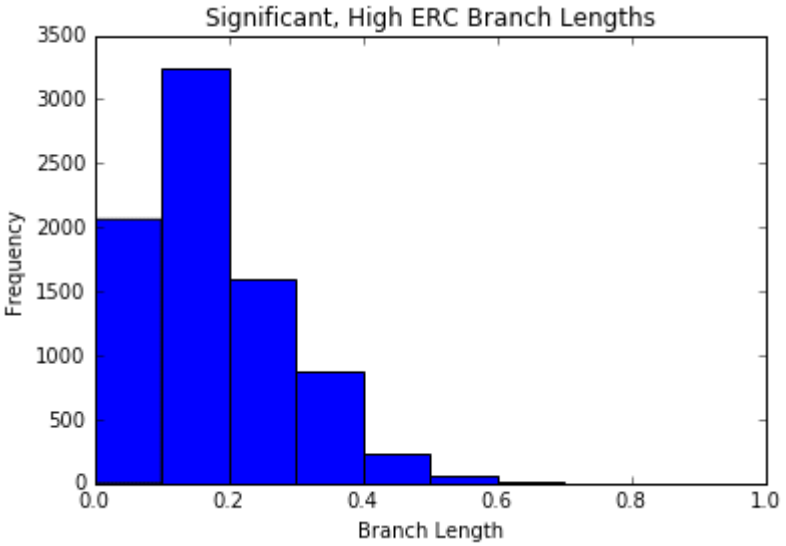
Just domain

Thursday, April 20, 2017 1:19 PM

Mean = 0.1955



Mean = 0.1933



Interpretation 2

Monday, April 24, 2017 12:05 PM

When evaluated for each individual protein or domain, the significant high ERC branch lengths are not noticeably shorter than the average branch length for the whole proteome.

When the domains are grouped by domain family, the ERC between the members of each pair domain families is averaged. Those domains with a high ERC value between their domain families show shorter branch lengths than the average for the whole proteome.

This could possibly because the domain family grouping is so vague that only a certain type of signal makes it through when looking for significant, high ERC. The "conserved proteins" signal is the only one that makes it through, while the "functionally related signal" gets muddled due to the broad groupings within domain families. Not sure

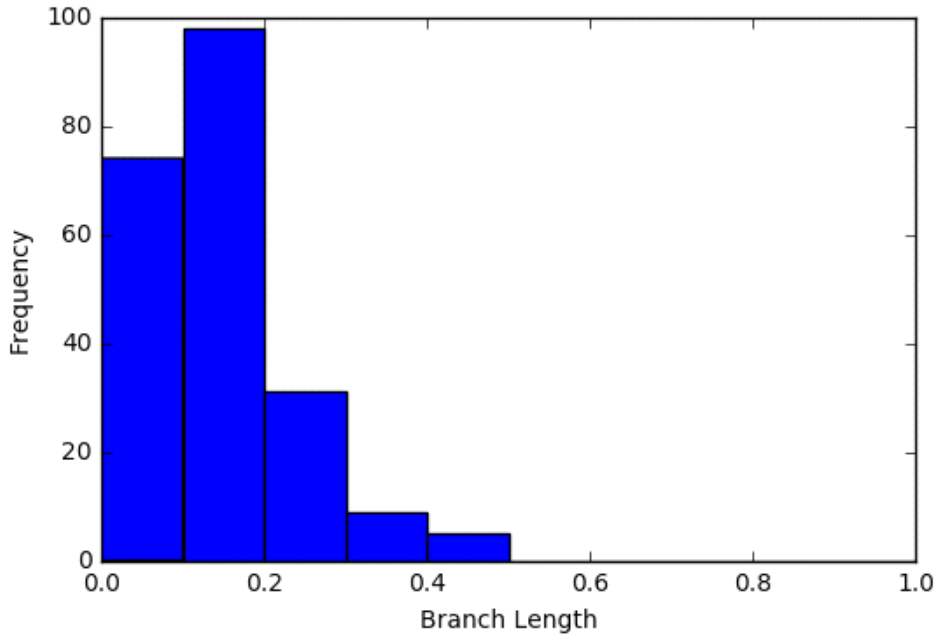
One difference between the analysis done on individual domains/proteins vs domain families is that the bonferroni correction was applied to the domain families, whereas the individual domains/proteins were given the largest correction I could manage to provide in python (

Check erc between domain family high erc vs low erc with distributions, branch lengths (see if one group is pulling it toward low branch length).

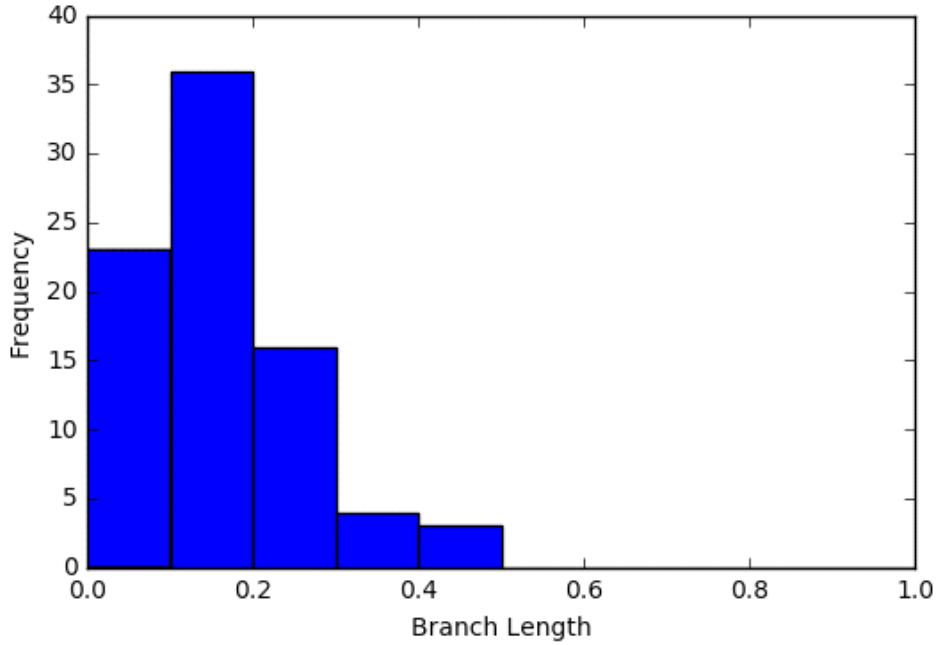
Compare ERC vs branch length

Monday, May 8, 2017 11:39 AM

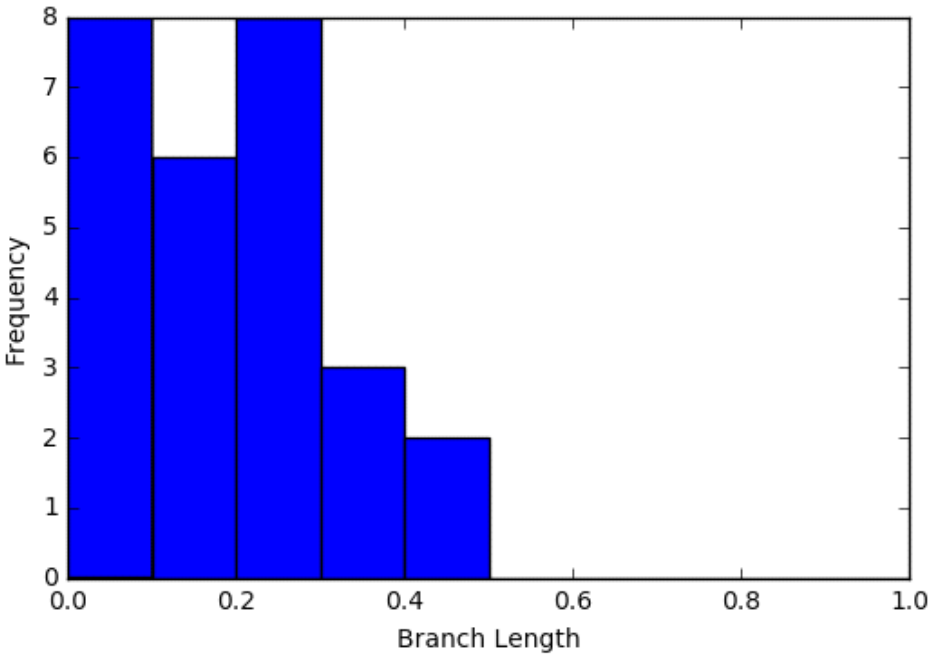
ERC > 0.2 ; mean = 0.1439



ERC > 0.3; mean = 0.1610



ERC > 0.4 ; mean = 0.188


Frequency
Branch Length
0
1
2
3
4
5
6
7
8
0.0
0.2
0.4
0.6
0.8
1.0