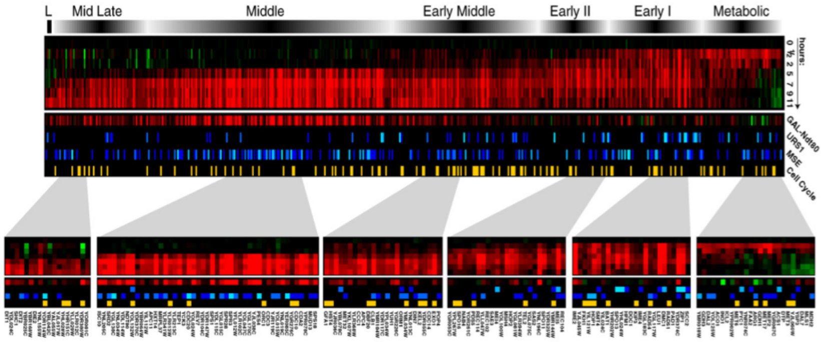
Bioinformatics and Functional Genomics wrapup

Biol4230 Thurs, April 26, 2018
Bill Pearson wrp@virginia.edu 4-2818 Pinn 6-057

Things not covered I:

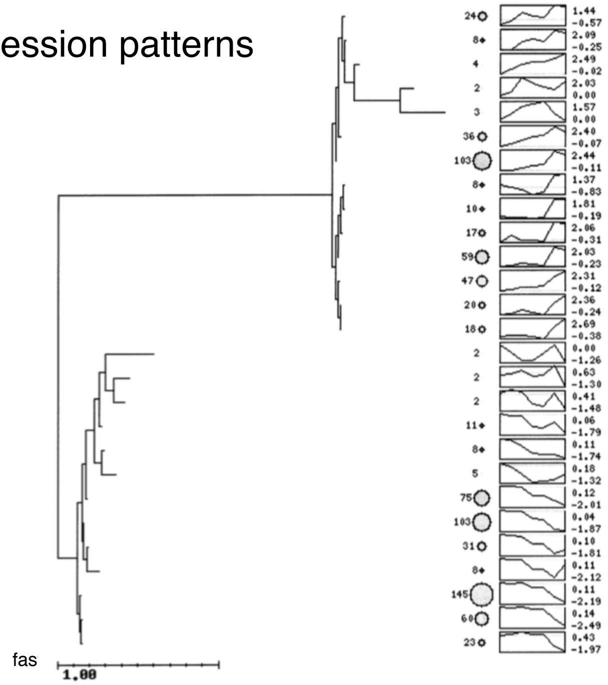
- Clustering and heat-maps
  - Principal Components Analysis revisited
  - Clustering strategies: k-means, hierarchical
    - when are the clusters "real"
- Function prediction/phenotype prediction
  - what does "function" mean? (trypsin vs chymotrypsin)
  - homologous proteins (usually) have similar functions – all function prediction is homology based
  - close homologs are more likely to have similar functions (but exceptions)

Yeast genes induced during sporulation



Chu, S. *et al. Science* **282,** 699–705 (1998).

Clustering of expression patterns



Clustering breast tumors by gene expression

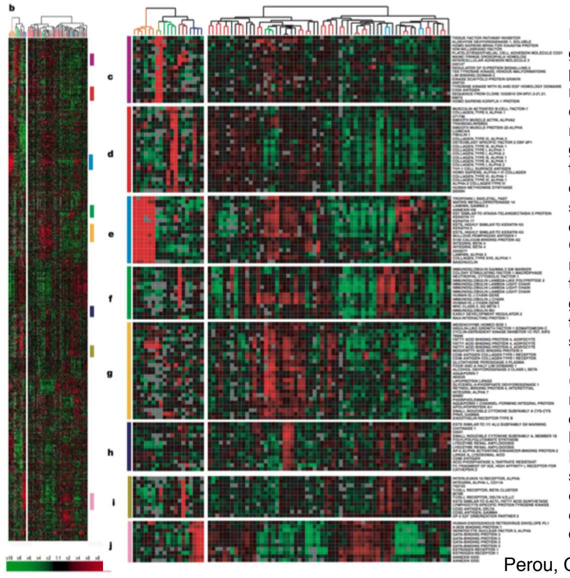


Figure 1 Variation in expression of 1,753 genes in 84 experimental samples. Data are presented in a matrix format: each row represents a single gene, and each column an experimental sample. In each sample, the ratio of the abundance of transcripts of each gene to the median abundance of the gene's transcript among all the cell lines (left panel), or to its median abundance across all tissue samples (right panel), is represented by the colour of the corresponding cell in the matrix. Green squares, transcript levels below the median; black squares, transcript levels equal to the median; red squares, transcript levels greater than the median; grey squares, technically inadequate or missing data. Colour saturation reflects the magnitude of the ratio relative to the median for each set of samples (see scale, bottom left). b, Scaled-down representation of the 1,753-gene cluster diagram; coloured bars to the right identify the locations of the inserts displayed in c-j. c, Endothelial cell gene expression cluster; d, stromal/fibroblast cluster; e, breast basal epithelial cluster; f, B-cell cluster; g, adipose-enriched/normal breast; h, macrophage; i, T-cell; j, breast luminal epithelial cell.

Perou, C. M. *et al. Nature* **406,** 747–752 (2000).

Clustering breast tumors by gene expression

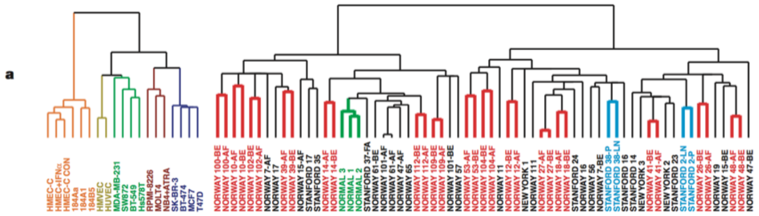


Figure 1 Variation in expression of 1,753 genes in 84 experimental samples. … a, Dendrogram representing similarities in the expression patterns between experimental samples. All `before and after' chemotherapy pairs that were clustered on terminal branches are highlighted in red; the two primary tumour/lymph node metastasis pairs in light blue; the three clustered normal breast samples in light green. Branches representing the four breast luminal epithelial cell lines are shown in dark blue; breast basal epithelial cell lines in orange, the endothelial cell lines in dark yellow, the mesynchemal-like cell lines in dark green, and the lymphocyte-derived cell lines in brown.

Perou, C. M. *et al. Nature* **406,** 747–752 (2000).

Clustering breast tumors by gene expression

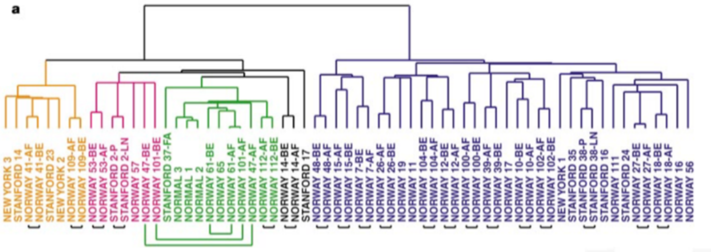
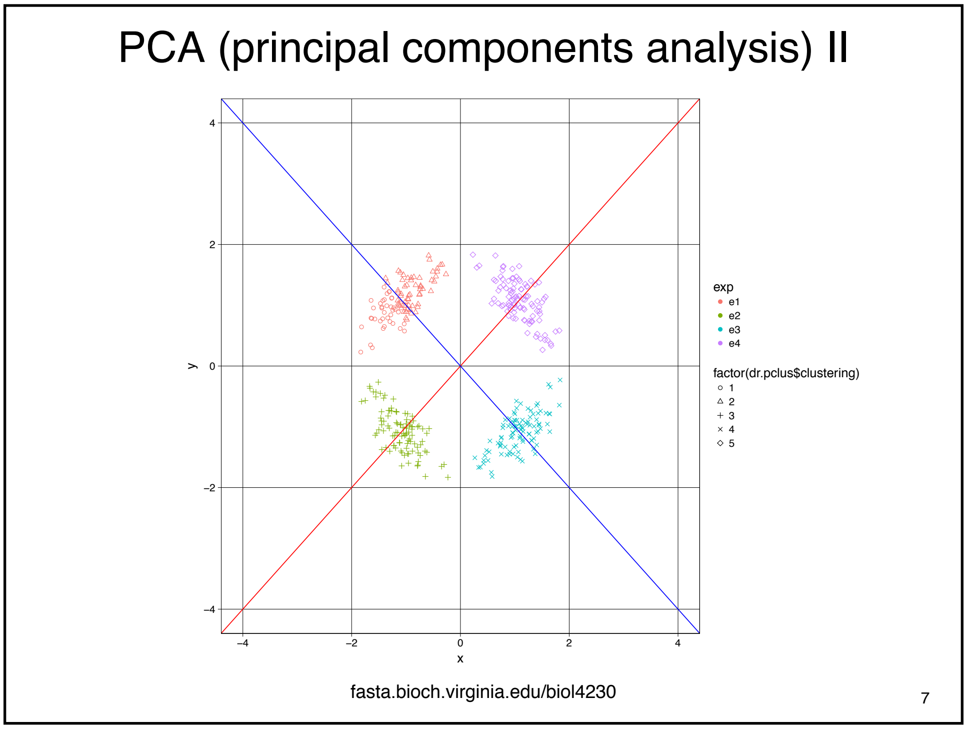
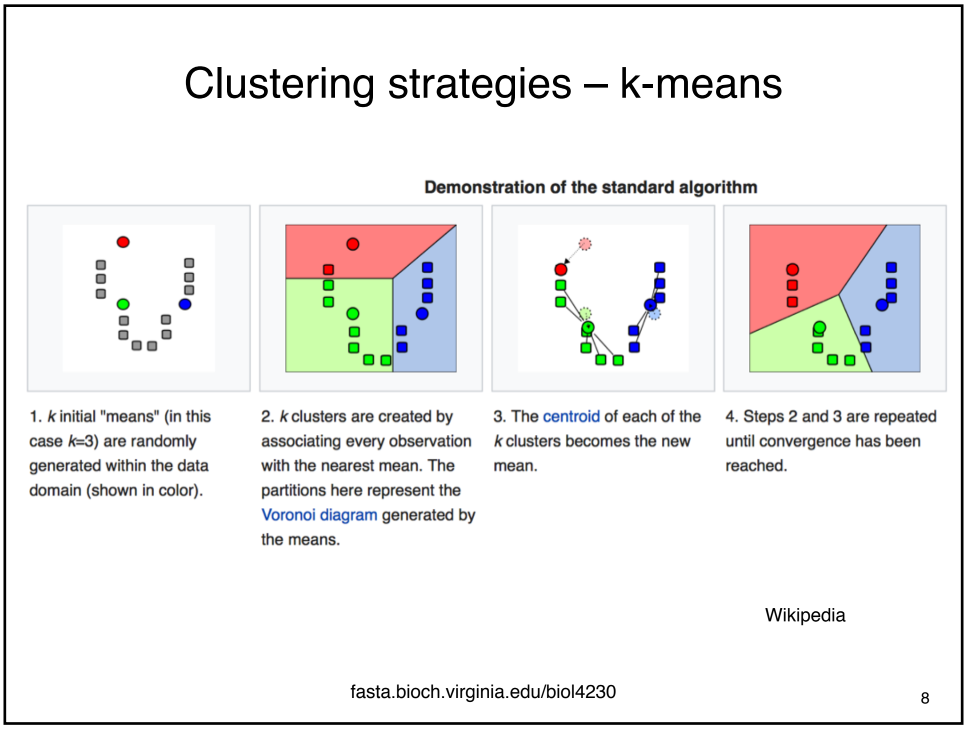


Figure 3 Cluster analysis using the `intrinsic' gene subset. Two large branches were apparent in the dendrogram, and within these large branches were smaller branches for which common biological themes could be inferred. Branches are coloured accordingly: basal-like, orange; Erb-B2+, pink; normal-breast-like, light green; and luminal epithelial/ER+, dark blue. a, Experimental sample associated cluster dendrogram. Small black bars beneath the dendrogram identify the 17 pairs that were matched by this hierarchical clustering; larger green bars identify the positions of the three pairs that were not matched by the clustering.

Perou, C. M. *et al. Nature* **406,** 747–752 (2000).



PCA (principal components analysis) II



Clustering strategies – k-means

Demonstration of the standard algorithm

1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

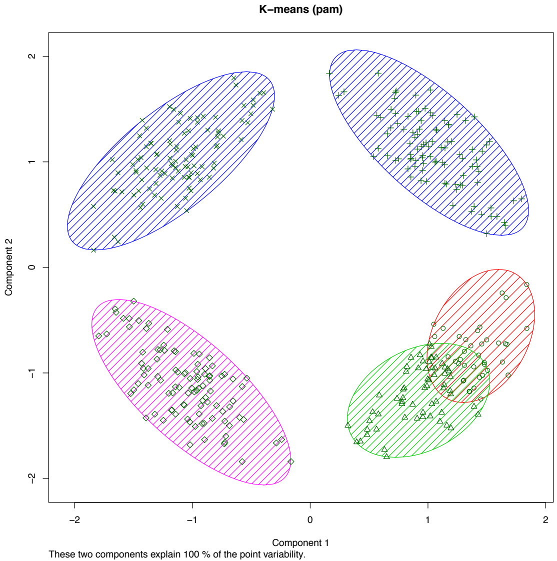
2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the *k* clusters becomes the new mean.

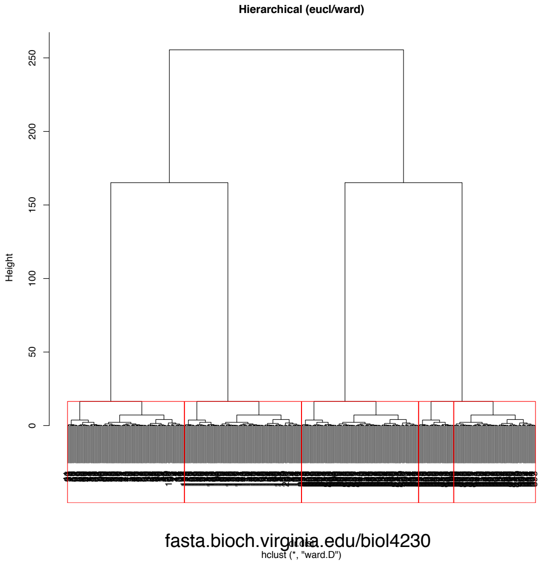
4. Steps 2 and 3 are repeated until convergence has been reached.

Wikipedia

Clustering strategies – k-means



Clustering strategies - hierarchical



From PCA to clustering

- PCA (principal components) reduces dimensionality – from 10,000 gene expression measurements to ? (10 or less)
- Clustering –
  - based on a distance measure (covariance)
  - many methods – k-means guarantee's k-clusters, right or wrong
  - hierarchical – are the relationships real?

Function and phenotype prediction

- what does "function" mean? (trypsin vs chymotrypsin)
- homologous proteins (usually) have similar functions – all function prediction is homology based
- close homologs are more likely to have similar functions (but exceptions)
- SIFT and Polyphen predict effect of mutations by building PSSMs

How to classify function:
E.C. (Enzyme Commission) numbers

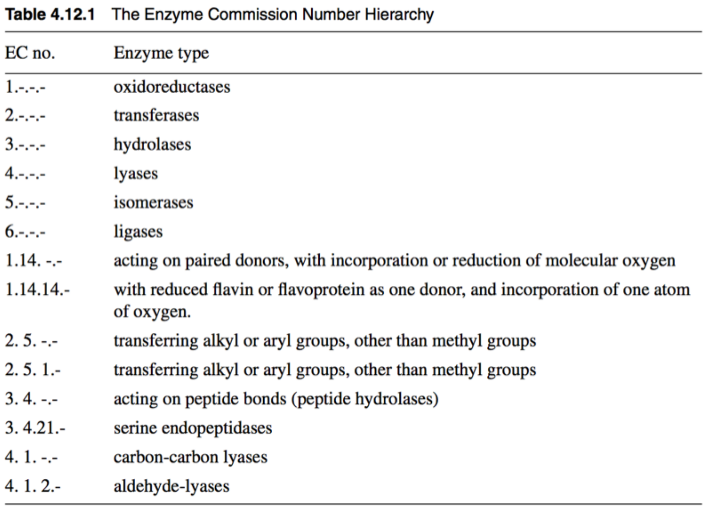
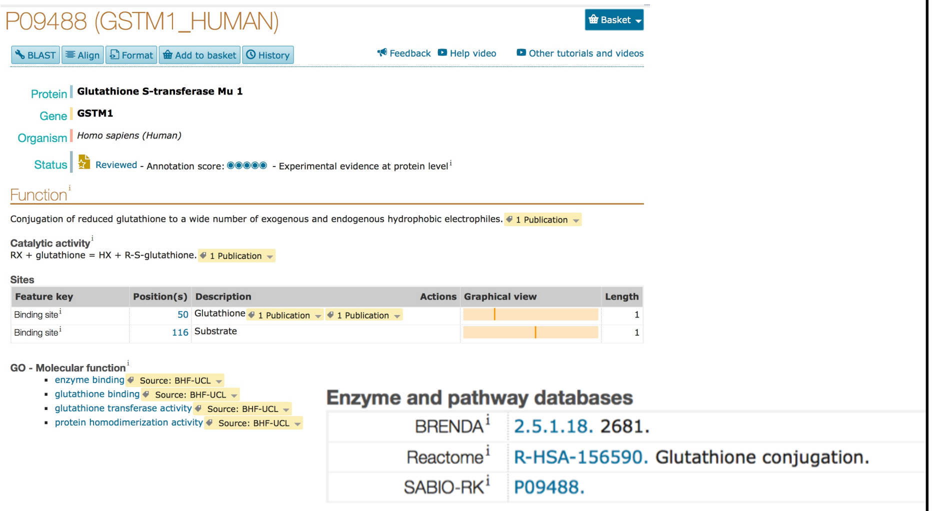


Table 4.12.1 The Enzyme Commission Number Hierarchy

EC no.	Enzyme type
1.-.-.-	oxidoreductases
2.-.-.-	transferases
3.-.-.-	hydrolases
4.-.-.-	lyases
5.-.-.-	isomerases
6.-.-.-	ligases
1.14. -.-	acting on paired donors, with incorporation or reduction of molecular oxygen
1.14.14.-	with reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen.
2. 5. -.-	transferring alkyl or aryl groups, other than methyl groups
2. 5. 1.-	transferring alkyl or aryl groups, other than methyl groups
3. 4. -.-	acting on peptide bonds (peptide hydrolases)
3. 4.21.-	serine endopeptidases
4. 1. -.-	carbon-carbon lyases
4. 1. 2.-	aldehyde-lyases

How to classify function:
E.C. (Enzyme Commission) numbers



P09488 (GSTM1_HUMAN)

Protein **Glutathione S-transferase Mu 1**
Gene **GSTM1**
Organism *Homo sapiens (Human)*
Status Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level[i]

Function[i]

Conjugation of reduced glutathione to a wide number of exogenous and endogenous hydrophobic electrophiles. 1 Publication

Catalytic activity[i]
RX + glutathione = HX + R-S-glutathione. 1 Publication

Sites

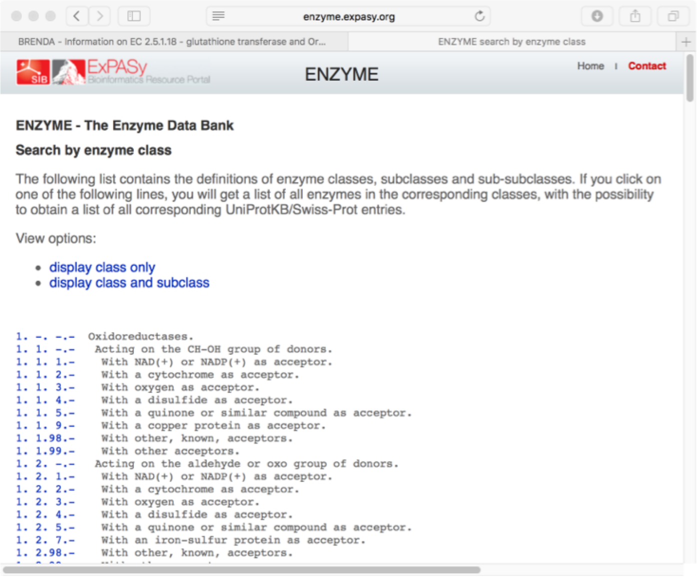
Feature key	Position(s)	Description	Actions	Graphical view	Length
Binding site[i]	50	Glutathione 1 Publication 1 Publication			1
Binding site[i]	116	Substrate			1

GO - Molecular function[i]
- enzyme binding Source: BHF-UCL
- glutathione binding Source: BHF-UCL
- glutathione transferase activity Source: BHF-UCL
- protein homodimerization activity Source: BHF-UCL

Enzyme and pathway databases

BRENDA[i]	2.5.1.18. 2681.
Reactome[i]	R-HSA-156590. Glutathione conjugation.
SABIO-RK[i]	P09488.

How to classify function: Enzyme/Expasy


enzyme.expasy.org
BRENDA - Information on EC 2.5.1.18 - glutathione transferase and Or...
ENZYME search by enzyme class
ExPASy
Bioinformatics Resource Portal
ENZYME
Home
Contact

ENZYME - The Enzyme Data Bank

Search by enzyme class

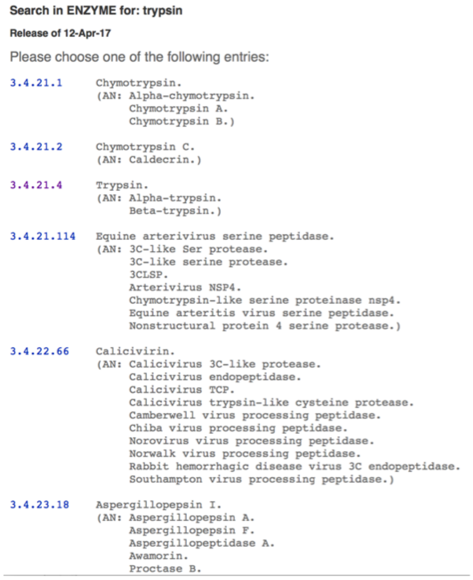
The following list contains the definitions of enzyme classes, subclasses and sub-subclasses. If you click on one of the following lines, you will get a list of all enzymes in the corresponding classes, with the possibility to obtain a list of all corresponding UniProtKB/Swiss-Prot entries.

View options:

- display class only
- display class and subclass

```
1. -. -.-   Oxidoreductases.
1. 1. -.-    Acting on the CH-OH group of donors.
1. 1. 1.-     With NAD(+) or NADP(+) as acceptor.
1. 1. 2.-     With a cytochrome as acceptor.
1. 1. 3.-     With oxygen as acceptor.
1. 1. 4.-     With a disulfide as acceptor.
1. 1. 5.-     With a quinone or similar compound as acceptor.
1. 1. 9.-     With a copper protein as acceptor.
1. 1.98.-     With other, known, acceptors.
1. 1.99.-     With other acceptors.
1. 2. -.-    Acting on the aldehyde or oxo group of donors.
1. 2. 1.-     With NAD(+) or NADP(+) as acceptor.
1. 2. 2.-     With a cytochrome as acceptor.
1. 2. 3.-     With oxygen as acceptor.
1. 2. 4.-     With a disulfide as acceptor.
1. 2. 5.-     With a quinone or similar compound as acceptor.
1. 2. 7.-     With an iron-sulfur protein as acceptor.
1. 2.98.-     With other, known, acceptors.
```

How to classify function: Enzyme/Expasy



Search in ENZYME for: trypsin

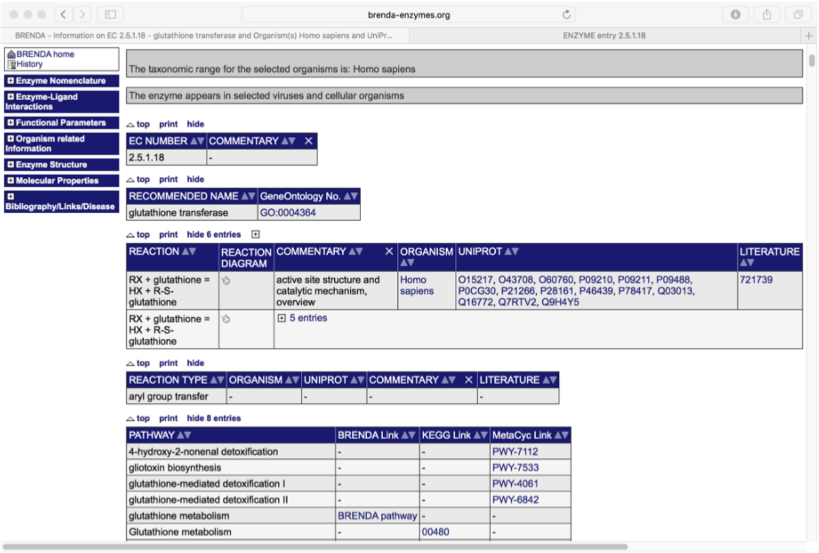
Release of 12-Apr-17

Please choose one of the following entries:

```
3.4.21.1     Chymotrypsin.
             (AN: Alpha-chymotrypsin.
                  Chymotrypsin A.
                  Chymotrypsin B.)

3.4.21.2     Chymotrypsin C.
             (AN: Caldecrin.)

3.4.21.4     Trypsin.
             (AN: Alpha-trypsin.
                  Beta-trypsin.)

3.4.21.114   Equine arterivirus serine peptidase.
             (AN: 3C-like Ser protease.
                  3C-like serine protease.
                  3CLSP.
                  Arterivirus NSP4.
                  Chymotrypsin-like serine proteinase nsp4.
                  Equine arteritis virus serine peptidase.
                  Nonstructural protein 4 serine protease.)

3.4.22.66    Calicivirin.
             (AN: Calicivirus 3C-like protease.
                  Calicivirus endopeptidase.
                  Calicivirus TCP.
                  Calicivirus trypsin-like cysteine protease.
                  Camberwell virus processing peptidase.
                  Chiba virus processing peptidase.
                  Norovirus virus processing peptidase.
                  Norwalk virus processing peptidase.
                  Rabbit hemorrhagic disease virus 3C endopeptidase.
                  Southampton virus processing peptidase.)

3.4.23.18    Aspergillopepsin I.
             (AN: Aspergillopepsin A.
                  Aspergillopepsin F.
                  Aspergillopeptidase A.
                  Awamorin.
                  Proctase B.
```

Different levels of the E.C. hierarchy do not consistently indicate different functional differences.

How to classify function: Brenda


brenda-enzymes.org
BRENDA - Information on EC 2.5.1.18 - glutathione transferase and Organism(s) Homo sapiens and UniPr...
ENZYME entry 2.5.1.18
BRENDA home
History
Enzyme Nomenclature
Enzyme-Ligand Interactions
Functional Parameters
Organism related Information
Enzyme Structure
Molecular Properties
Bibliography/Links/Disease

The taxonomic range for the selected organisms is: Homo sapiens

The enzyme appears in selected viruses and cellular organisms

top print hide

EC NUMBER	COMMENTARY
2.5.1.18	-

top print hide

RECOMMENDED NAME	GeneOntology No.
glutathione transferase	GO:0004364

top print hide 6 entries

REACTION	REACTION DIAGRAM	COMMENTARY	ORGANISM	UNIPROT	LITERATURE
RX + glutathione = HX + R-S-glutathione		active site structure and catalytic mechanism, overview	Homo sapiens	O15217, O43708, O60760, P09210, P09211, P09488, P0CG30, P21266, P28161, P46439, P78417, Q03013, Q16772, Q7RTV2, Q9H4Y5	721739
RX + glutathione = HX + R-S-glutathione		5 entries			

top print hide

REACTION TYPE	ORGANISM	UNIPROT	COMMENTARY	LITERATURE
aryl group transfer	-	-	-	-

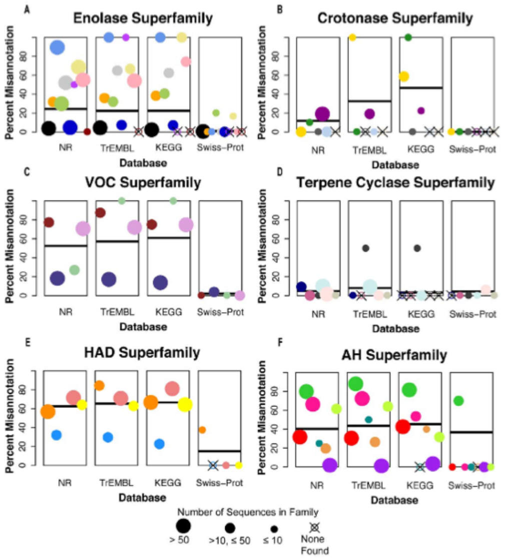
top print hide 8 entries

PATHWAY	BRENDA Link	KEGG Link	MetaCyc Link
4-hydroxy-2-nonenal detoxification	-	-	PWY-7112
gliotoxin biosynthesis	-	-	PWY-7533
glutathione-mediated detoxification I	-	-	PWY-4061
glutathione-mediated detoxification II	-	-	PWY-6842
glutathione metabolism	BRENDA pathway	-	-
Glutathione metabolism	-	00480	-

fasta.bioch.virginia.edu/biol4230

17

Inference of Function from Homology



- SwissProt is very accurate
- NR and Trembl make no claim to functional accuracy (all databases are not equal; bigger ≠ better)

A. M. Schnoes, S. D. Brown, Igor Dodevski, P. C. Babbitt (2009) Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies PLOS Comput. Biol. 5:e1000605

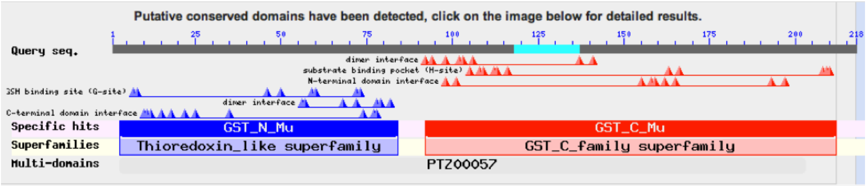
Inferring Function – Critical Information

- Homologous proteins *always* have similar structures, but need not have similar functions
- BLAST and FASTA obscure information required to infer function
- Even with appropriate information, inferring function is challenging

- Homology – E() value
- Alignment location
- Catalytic activity of homologs
- State of active site residues

Currently, similarity searching programs focus on homology, and fail to present available functional annotation

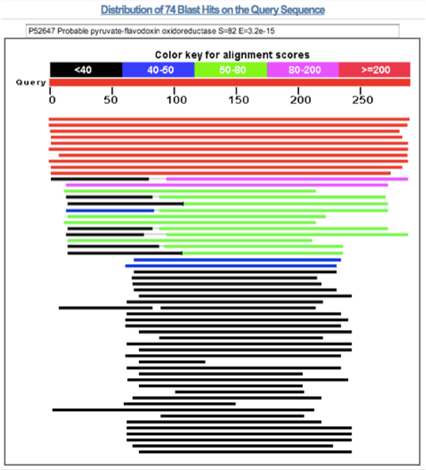
Conventional sequence alignments do not show functional sites (and even if they did, we would not look)

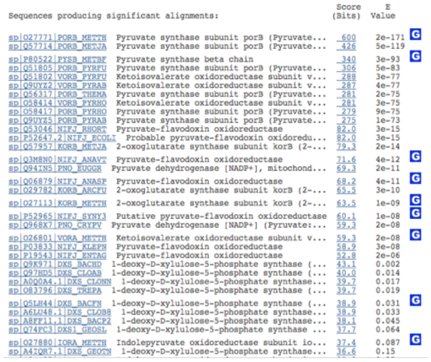


Putative conserved domains have been detected, click on the image below for detailed results.

- Shows conserved domains, and annotated residues
- Does not show state (or even coordinate) of annotated residues in query or homologs

Search results obscure functional information





```
                                                                      Score    E
Sequences producing significant alignments:                          (Bits)  Value

sp|O27771|PORB_METTH  Pyruvate synthase subunit porB (Pyruvate...    600    2e-171  G
sp|Q57714|PORB_METJA  Pyruvate synthase subunit porB (Pyruvate...    426    5e-119

sp|P80522|PYSB_METBF  Pyruvate synthase beta chain                   340    3e-93   G
sp|Q51805|PORB_PYRFU  Pyruvate synthase subunit porB (Pyruvate...    306    5e-83
sp|Q51802|VORB_PYRFU  Ketoisovalerate oxidoreductase subunit v...    288    3e-77
sp|Q9UYZ2|VORB_PYRAB  Ketoisovalerate oxidoreductase subunit v...    287    4e-77
sp|Q56317|PORB_THEMA  Pyruvate synthase subunit porB (Pyruvate...    281    2e-75
sp|O58414|VORB_PYRHO  Ketoisovalerate oxidoreductase subunit v...    281    3e-75
sp|O58417|PORB_PYRHO  Pyruvate synthase subunit porB (Pyruvate...    279    9e-75
sp|Q9UYZ5|PORB_PYRAB  Pyruvate synthase subunit porB (Pyruvate...    275    1e-73
sp|Q53046|NIFJ_RHORT  Pyruvate-flavodoxin oxidoreductase             82.0   3e-15
sp|P52647.2|NIFJ_ECOLI  Probable pyruvate-flavodoxin oxidoredu...    82.0   3e-15
sp|Q57957|KORB_METJA  2-oxoglutarate synthase subunit korB (2-...    79.3   2e-14

sp|Q3M8N0|NIFJ_ANAVT  Pyruvate-flavodoxin oxidoreductase             71.6   4e-12   G
sp|Q94IN5|PNO_EUGGR  Pyruvate dehydrogenase [NADP+], mitochond...    69.3   2e-11

sp|Q06879|NIFJ_ANASP  Pyruvate-flavodoxin oxidoreductase             68.2   4e-11   G
sp|O29782|KORB_ARCFU  2-oxoglutarate synthase subunit korB (2-...    65.5   3e-10

sp|O27113|KORB_METTH  2-oxoglutarate synthase subunit korB (2-...    63.5   1e-09   G

sp|P52965|NIFJ_SYNY3  Putative pyruvate-flavodoxin oxidoreductase    60.1   1e-08   G
sp|Q968X7|PNO_CRYPV  Pyruvate dehydrogenase [NADP+] (Pyruvate:...    59.3   2e-08

sp|O26801|VORA_METTH  Ketoisovalerate oxidoreductase subunit v...    59.3   2e-08   G
sp|P03833|NIFJ_KLEPN  Pyruvate-flavodoxin oxidoreductase             58.9   3e-08
sp|P19543|NIFJ_ENTAG  Pyruvate-flavodoxin oxidoreductase             52.8   2e-06
sp|Q9K971|DXS_BACHD  1-deoxy-D-xylulose-5-phosphate synthase (...    43.1   0.002
sp|Q97HD5|DXS_CLOAB  1-deoxy-D-xylulose-5-phosphate synthase (...    40.0   0.014
sp|A0Q0A4.1|DXS_CLONN  1-deoxy-D-xylulose-5-phosphate synthase...    39.7   0.017
sp|O83796|DXS_TREPA  1-deoxy-D-xylulose-5-phosphate synthase (...    39.7   0.019

sp|Q5LH44|DXS_BACFN  1-deoxy-D-xylulose-5-phosphate synthase (...    38.9   0.031   G
sp|A6LU48.1|DXS_CLOB8  1-deoxy-D-xylulose-5-phosphate synthase...    38.9   0.033
sp|A8FF11.1|DXS_BACP2  1-deoxy-D-xylulose-5-phosphate synthase...    38.1   0.045
sp|Q74FC3|DXS1_GEOSL  1-deoxy-D-xylulose-5-phosphate synthase ...    37.7   0.064

sp|O27880|IORA_METTH  Indolepyruvate oxidoreductase subunit io...    37.4   0.087   G
sp|A4IQR7.1|DXS_GEOTN  1-deoxy-D-xylulose-5-phosphate synthase...    36.6   0.15
```

Similarity Search Results – NCBI/BLAST

21

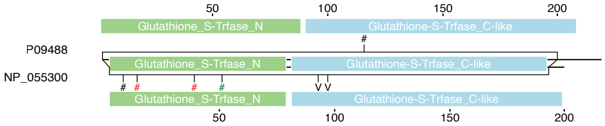
Annotations from Uniprot

```
ID   GSTM1_HUMAN             Reviewed;         218 AA.
DT   28-NOV-2012, entry version 148.
DE   RecName: Full=Glutathione S-transferase Mu 1;
GN   Name=GSTM1; Synonyms=GST1;
...
FT   DOMAIN        2     88       GST N-terminal.
FT   DOMAIN       90    208       GST C-terminal.
FT   REGION        7      8       Glutathione binding.
FT   REGION       46     50       Glutathione binding.
FT   REGION       59     60       Glutathione binding.
FT   REGION       72     73       Glutathione binding.
FT   BINDING     116    116       Substrate.
FT   MOD_RES      23     23       Phosphotyrosine (By similarity).
FT   MOD_RES      33     33       Phosphotyrosine (By similarity).
FT   MOD_RES      34     34       Phosphothreonine (By similarity).
FT   VAR_SEQ     153    189       Missing (in isoform 2).
FT   VARIANT     173    173       K -> N (in allele GSTM1B; dbSNP:rs1065411).
FT   VARIANT     210    210       S -> T (in dbSNP:rs449856).
FT   MUTAGEN       7      7       Y->F: Reduces catalytic activity 100-fold.
FT   MUTAGEN     108    108       H->Q: Reduces catalytic activity by half.
FT   MUTAGEN     108    108       H->S: Changes the properties of the enzyme.
FT   MUTAGEN     109    109       M->I: Reduces catalytic activity by half.
FT   MUTAGEN     116    116       Y->A: Reduces catalytic activity 10-fold.
FT   MUTAGEN     116    116       Y->F: Slight increase of catalytic activity
```

Alignments with Annotations

FASTA-36.3.6 output:

```
>>sp|P09488|GSTM1_HUMAN                                                (218 aa) vs
>>ref|NP_055300.1| prostaglandin-D synthase [Homo sapiens]             (199 aa)
 Site:# : 7Y=8Y : BINDING: Glutathione.
 Site:# : 13L<14R : BINDING: Glutathione.
 Site:# : 46W=39W : BINDING: Glutathione.
 Site:# : 52K=45K : BINDING: Glutathione (By similarity).
qSite:# : 116Y=109Y : BINDING: Substrate.
 Site:* : 136K=128K : MOD_RES: N6-acetyllysine.
qVariant: 108Q>101R : H101Q : Mutagen: Reduces catalytic activity by half.
 Variant: 112G<105V : I105V : in allele GSTP1*B and allele GSTP1*C; dbSNP:rs1695.
 Variant: 173K<169D : G169D : in dbSNP:rs41462048.
qVariant: 173Nz169D : K169N : in allele GSTM1B; dbSNP:rs1065411.
qRegion: 2-88:3-81 : score=83; bits=37.2; Id=0.287; Q=65.5 :  Glutathione_S-Trfase_N
qRegion: 90-208:83-204 : score=158; bits=66.0; Id=0.285; Q=151.9 :  Glutathione-S-Trfase_C-like
 Region: 2-88:3-81 : score=83; bits=37.2; Id=0.287; Q=65.5 :  Glutathione_S-Trfase_N
 Region: 90-208:83-204 : score=156; bits=65.2; Id=0.285; Q=149.6 :  Glutathione-S-Trfase_C-like
 s-w opt: 242  Z-score: 492.1  bits: 98.1 E(35695): 4.8e-21
Smith-Waterman score: 242; 28.4% identity (63.5% similar) in 211 aa overlap (2-208:3-204)
```

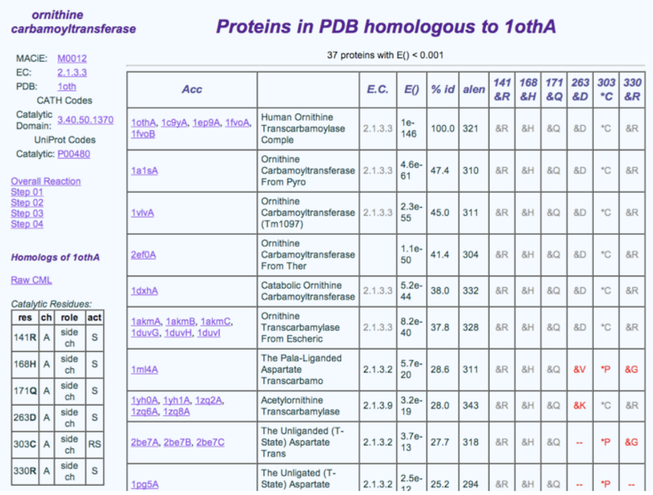


Capturing variation, functional sites, and domain similarity with FASTA/SSEARCH

Annotations extracted from uniprot_sprot.dat features:

```
>sp|P09488|GSTM1_HUMAN
2        -        88       DOMAIN: GST N-terminal.
7        V        F        Mutagen: Reduces catalytic activity 100- fold.
23       *        -        MOD_RES: Phosphotyrosine (By similarity).
33       *        -        MOD_RES: Phosphotyrosine (By similarity).
34       *        -        MOD_RES: Phosphothreonine (By similarity).
90       -        208      DOMAIN: GST C-terminal.
108      V        S        Mutagen: Changes the prop. of the enzyme toward
some subs.
108      V        Q        Mutagen: Reduces catalytic activity by half.
109      V        I        Mutagen: Reduces catalytic activity by half.
116      #        -        BINDING: Substrate.
116      V        A        Mutagen: Reduces catalytic activity 10-fold.
116      V        F        Mutagen: Slight increase of catalytic activity.
173      V        N        in allele GSTM1B; dbSNP:rs1065411.
210      V        T        in dbSNP:rs449856.
```

Highlighting Active Site state (MACIE)



ornithine carbamoyltransferase

Proteins in PDB homologous to 1othA

MACiE: M0012
EC: 2.1.3.3
PDB: 1oth

CATH Codes
Catalytic Domain: 3.40.50.1370

UniProt Codes
Catalytic: P00480

Overall Reaction
Step 01
Step 02
Step 03
Step 04

Homologs of 1othA

Raw CML

Catalytic Residues:

res	ch	role	act
141R	A	side ch	S
168H	A	side ch	S
171Q	A	side ch	S
263D	A	side ch	S
303C	A	side ch	RS
330R	A	side ch	S

37 proteins with E() < 0.001

Acc		E.C.	E()	% id	alen	141 &R	168 &H	171 &Q	263 &D	303 *C	330 &R
1othA, 1c9yA, 1ep9A, 1fvoA, 1fvoB	Human Ornithine Transcarbamoylase Comple	2.1.3.3	1e-146	100.0	321	&R	&H	&Q	&D	*C	&R
1a1sA	Ornithine Carbamoyltransferase From Pyro	2.1.3.3	4.6e-61	47.4	310	&R	&H	&Q	&D	*C	&R
1vlvA	Ornithine Carbamoyltransferase (Tm1097)	2.1.3.3	2.3e-55	45.0	311	&R	&H	&Q	&D	*C	&R
2ef0A	Ornithine Carbamoyltransferase From Ther		1.1e-50	41.4	304	&R	&H	&Q	&D	*C	&R
1dxhA	Catabolic Ornithine Carbamoyltransferase	2.1.3.3	5.2e-44	38.0	332	&R	&H	&Q	&D	*C	&R
1akmA, 1akmB, 1akmC, 1duvG, 1duvH, 1duvI	Ornithine Transcarbamylase From Escheric	2.1.3.3	8.2e-40	37.8	328	&R	&H	&Q	&D	*C	&R
1ml4A	The Pala-Liganded Aspartate Transcarbamo	2.1.3.2	5.7e-20	28.6	311	&R	&H	&Q	&V	*P	&G
1yh0A, 1yh1A, 1zq2A, 1zq6A, 1zq8A	Acetylornithine Transcarbamylase	2.1.3.9	3.2e-19	28.0	343	&R	&H	&Q	&K	*C	&R
2be7A, 2be7B, 2be7C	The Unliganded (T-State) Aspartate Trans	2.1.3.2	3.7e-13	27.7	318	&R	&H	&Q	--	*P	&G
1pg5A	The Unligated (T-State) Aspartate	2.1.3.2	2.5e-12	25.2	294	&R	&H	&Q	--	*P	--

Holliday et al (2012) NAR

Highlighting Active Site state (MACIE)

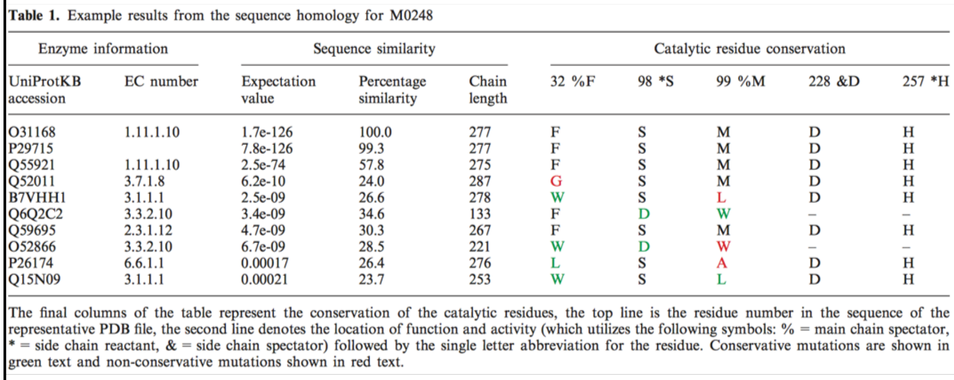


Table 1. Example results from the sequence homology for M0248

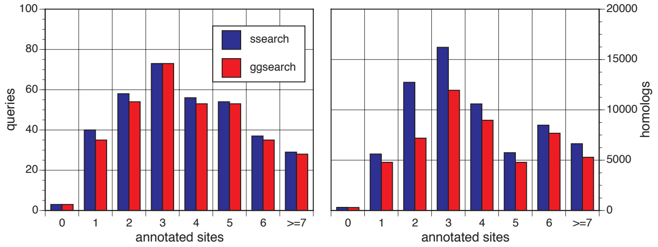
Enzyme information		Sequence similarity			Catalytic residue conservation				
UniProtKB accession	EC number	Expectation value	Percentage similarity	Chain length	32 %F	98 *S	99 %M	228 &D	257 *H
O31168	1.11.1.10	1.7e-126	100.0	277	F	S	M	D	H
P29715		7.8e-126	99.3	277	F	S	M	D	H
Q55921	1.11.1.10	2.5e-74	57.8	275	F	S	M	D	H
Q52011	3.7.1.8	6.2e-10	24.0	287	G	S	M	D	H
B7VHH1	3.1.1.1	2.5e-09	26.6	278	W	S	L	D	H
Q6Q2C2	3.3.2.10	3.4e-09	34.6	133	F	D	W	–	–
Q59695	2.3.1.12	4.7e-09	30.3	267	F	S	M	D	H
O52866	3.3.2.10	6.7e-09	28.5	221	W	D	W	–	–
P26174	6.6.1.1	0.00017	26.4	276	L	S	A	D	H
Q15N09	3.1.1.1	0.00021	23.7	253	W	S	L	D	H

The final columns of the table represent the conservation of the catalytic residues, the top line is the residue number in the sequence of the representative PDB file, the second line denotes the location of function and activity (which utilizes the following symbols: % = main chain spectator, * = side chain reactant, & = side chain spectator) followed by the single letter abbreviation for the residue. Conservative mutations are shown in green text and non-conservative mutations shown in red text.

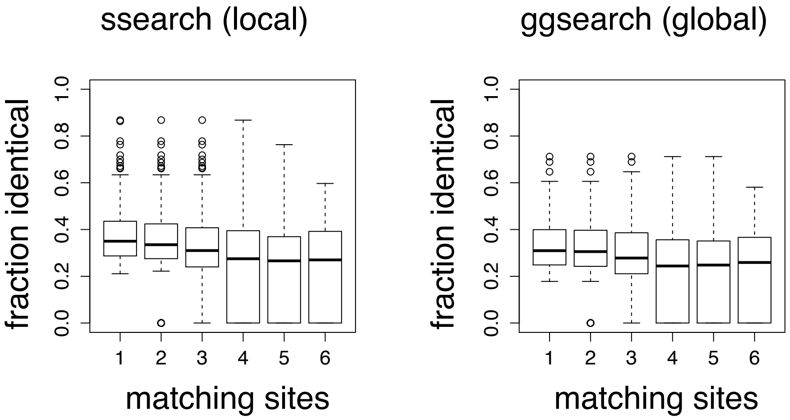
Holliday et al (2012) NAR

Active site conservation improves function prediction

- Search with ~400 proteins of known structure, function (E.C. number), sites from MACiE
- Find locally (ssearch36) or globally (ggsearch36) similar homologs
- Very few proteins with >50% global identity with different EC3 numbers
- Matching all annotated sites improves prediction sensitivity

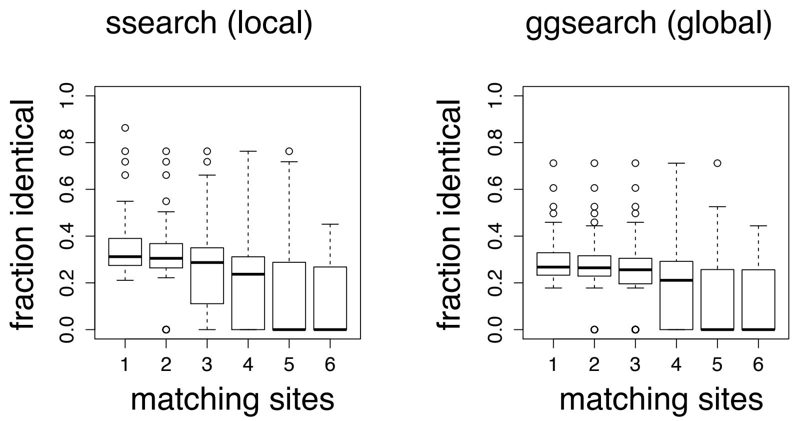


Annotations improve sensitivity (percent identity of first different EC4)


ssearch (local)

ggsearch (global)

Annotations improve sensitivity (percent identity of first different EC3)



Predicting mutation phenotype – SIFT and Polyphen

- SIFT – Sort Intolerant From Tolerant substitutions
  - Find protein homologs (PSI-BLAST)
  - Build PSSM
  - Use PSSM, rather than BLOSUM62, to predict phenotype (tolerated/not-tolerated)
- PolyPhen-2
  - Find homologs, multiple alignment
  - Find homologous structures
  - Combine PSSM, identity, Pfam domains, residue volume, etc…

Function follows conservation

Conservation

Function

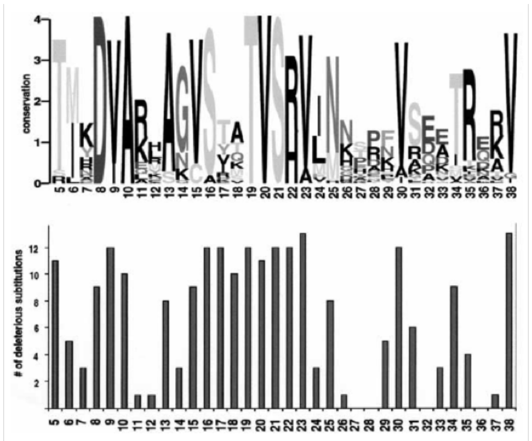


Figure 1 Sequence conservation corresponds to intolerant positions. (*Top*) Sequence logo representation (Schneider and Stephens 1990) of the LacI multiple alignment for positions 5–38, a region involved in binding DNA. At each position, the stack of letters indicates which amino acids appear in the alignment, and the total height of the stack is a measure of conservation. (*Bottom*) Number of substitutions deleterious to LacI function at the corresponding positions (Markiewicz et al. 1994; Suckow et al. 1996). Positions with high conservation, such as 19–23, do not tolerate substitutions. Positions with low conservation, such as 26–28, can tolerate most substitutions. Positions 17 and 18 appear diverse in the alignment but cannot tolerate most substitutions. The side chains of these residues are involved in DNA-specific recognition (Chuprina et al. 1993) that is not conserved among the paralogous sequences.

Ng and Henikoff, (2001) Gen. Res. 11:863

Position-Specific Scores
ATP Synthase, 4 iterations

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	bits/pos
BL62	Q	−1	1	0	0	−3	5	2	−2	0	−3	−2	1	0	−3	−1	0	−1	−2	−1	−2	0.70
46	Q	−2	−1	−2	−2	−4	6	0	1	0	−4	−3	−1	−2	−1	−3	−1	−2	6	4	−3	0.74
	%	0	0	0	0	0	54	0	12	0	0	0	0	0	0	0	0	0	13	20	0	
47	Q	−1	−1	3	3	−3	3	3	−2	3	−4	−4	−1	−3	−4	−2	2	−1	−4	−2	−3	0.51
	%	0	0	13	20	0	16	19	0	8	0	0	0	0	0	0	24	0	0	0	0	
56	Q	−2	−1	−2	−2	−3	5	2	−4	−1	4	−1	−1	−1	−2	−3	−2	−2	−3	−2	0	0.51
	%	0	0	0	0	0	46	13	0	0	41	0	0	0	0	0	0	0	0	0	0	
97	Q	−2	−1	0	−2	−4	4	0	−3	8	−4	−4	−1	−2	−3	−3	−1	−2	−3	0	−4	1.11
	%	0	0	0	0	0	35	0	0	65	0	0	0	0	0	0	0	0	0	0	0	
131	Q	3	−1	−1	−1	−2	5	2	−2	−1	−3	−3	0	−2	−4	−2	1	−1	−3	−3	−2	0.52
	%	44	0	0	0	0	36	11	0	0	0	0	0	0	0	0	9	0	0	0	0	
152	Q	−2	6	−1	−2	−4	4	0	−3	−1	−4	−3	1	−2	−4	−3	−1	−2	−4	−3	−3	1.00
	%	0	77	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
210	Q	−2	0	−1	−1	−4	7	1	−3	0	−4	−3	1	−1	−4	−2	−1	−2	−3	−2	−3	1.13
	%	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

SIFT (PSSMs) out performs BLOSUM62

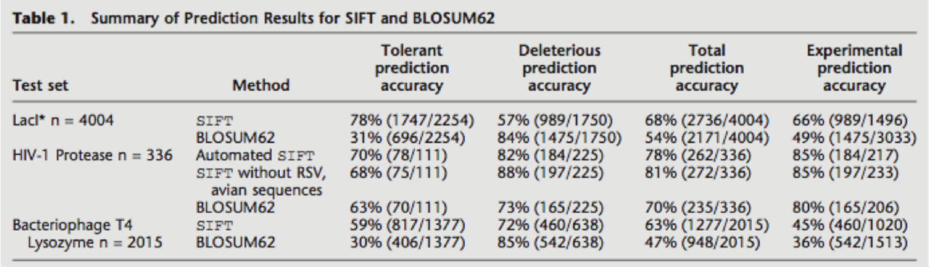
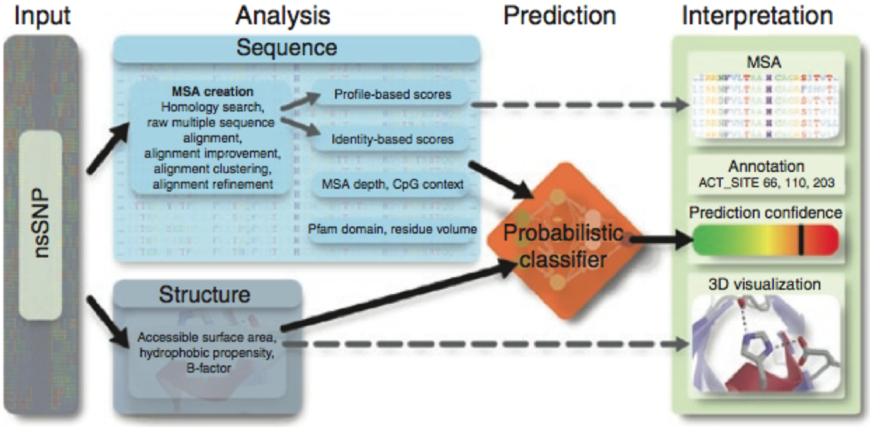


Table 1. Summary of Prediction Results for SIFT and BLOSUM62

Test set	Method	Tolerant prediction accuracy	Deleterious prediction accuracy	Total prediction accuracy	Experimental prediction accuracy
LacI* n = 4004	SIFT	78% (1747/2254)	57% (989/1750)	68% (2736/4004)	66% (989/1496)
	BLOSUM62	31% (696/2254)	84% (1475/1750)	54% (2171/4004)	49% (1475/3033)
HIV-1 Protease n = 336	Automated SIFT	70% (78/111)	82% (184/225)	78% (262/336)	85% (184/217)
	SIFT without RSV, avian sequences	68% (75/111)	88% (197/225)	81% (272/336)	85% (197/233)
	BLOSUM62	63% (70/111)	73% (165/225)	70% (235/336)	80% (165/206)
Bacteriophage T4 Lysozyme n = 2015	SIFT	59% (817/1377)	72% (460/638)	63% (1277/2015)	45% (460/1020)
	BLOSUM62	30% (406/1377)	85% (542/638)	47% (948/2015)	36% (542/1513)

Ng and Henikoff, (2001) Genome Res. 11:863

PolyPhen(2) – MSA, PSSM, structure, + ?



Input

Analysis

Prediction

Interpretation

nsSNP

Sequence

MSA creation
Homology search,
raw multiple sequence
alignment,
alignment improvement,
alignment clustering,
alignment refinement

Profile-based scores

Identity-based scores

MSA depth, CpG context

Pfam domain, residue volume

Structure

Accessible surface area,
hydrophobic propensity,
B-factor

Probabilistic classifier

MSA

Annotation
ACT_SITE 66, 110, 203

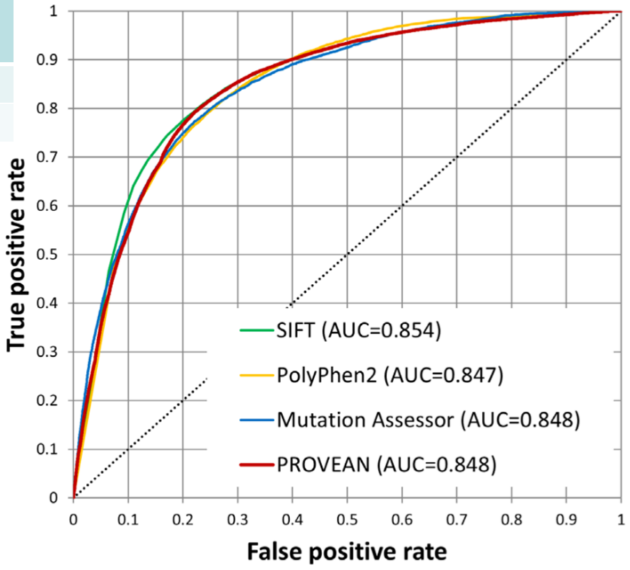
Prediction confidence

3D visualization

Adzhubei et al (2010) Nat. Methods 7:248

Evaluating prediction performance: ROC (receiver operator characteristic) curves

Predict: Expt	Yes	No
Yes	TP	FP
No	FN	TP



Choi, Y. et al (2012) *PLoS ONE* **7,** e46688 (2012).

SIFT has high sensitivity, but many false positives (low specificity)

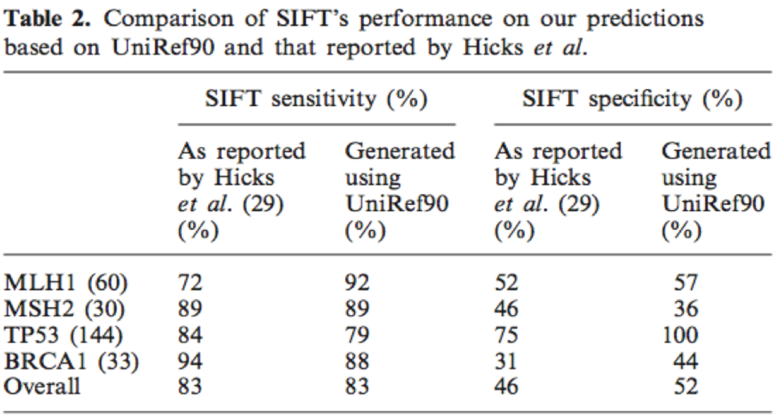
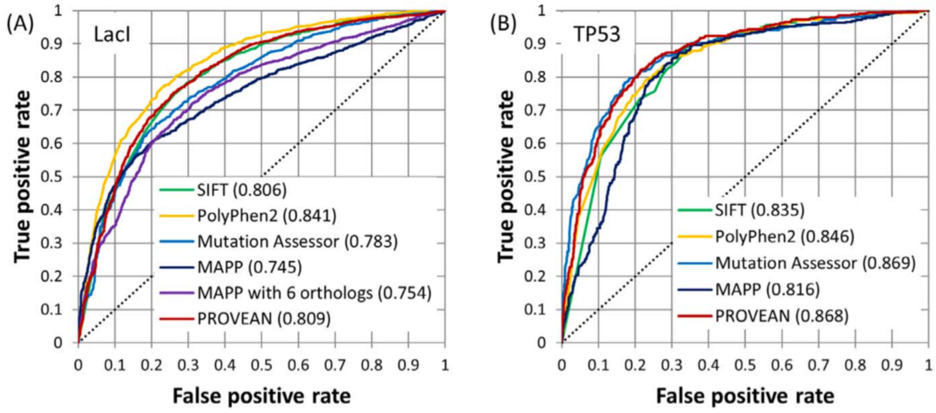


Table 2. Comparison of SIFT's performance on our predictions based on UniRef90 and that reported by Hicks *et al.*

	SIFT sensitivity (%)		SIFT specificity (%)	
	As reported by Hicks *et al.* (29) (%)	Generated using UniRef90 (%)	As reported by Hicks *et al.* (29) (%)	Generated using UniRef90 (%)
MLH1 (60)	72	92	52	57
MSH2 (30)	89	89	46	36
TP53 (144)	84	79	75	100
BRCA1 (33)	94	88	31	44
Overall	83	83	46	52

Sim et al. (2012) Nuc Acids Res 40:W:452

Evaluating prediction performance: slight differences for different proteins



Choi, Y. et al (2012) *PLoS ONE*
7, e46688 (2012).

Phenotype Prediction: SIFT/PolyPhen

- Traditional scoring matrices (BLOSUM62) make useful predictions about deleterious mutations
- Family-specific matrices (PSSMs) do better (SIFT)
- Including additional structural and domain information improves prediction slightly (PolyPhen2)
- All methods work as filters, but require confirmation