1

Bioinformatic Methods I

Week	Topic
1	NCBI/Blast I
2	**Blast II/Comparative Genomics**
3	Multiple Sequence Alignments
4	Phylogenetics
5	Selection Analysis
6	NGS Analysis / Metagenomics

2

Why identify similar sequences?

- Similarity is the primary predictor of homology.
- Homology is the primary computational predictor of function.
- Sequence alignments allow us to identify similar sequences:

```
Sequence 1: HEAGAWGHEE
Sequence 2: PAWHEAE

Sequence 1: HEAGAWGHE-E
              .  ++  ++  +
Sequence 2: --P-AW-HEAE
```


Overview

Substitution Matrices

Alignment Methods

- **Dot Matrix**
- Dynamic Programming
  - Global Alignment
  - Local Alignment
- Hidden Markov Models
- Heuristic Alignment - k-tuple
  - **BLAST**
  - FASTA

Evaluation of Significance

Comparative Genomics

Substitution matrices

Scoring systems
- model sequence change over evolutionary time

Realistic models of sequence evolution
- substitution biases
- mutational saturation

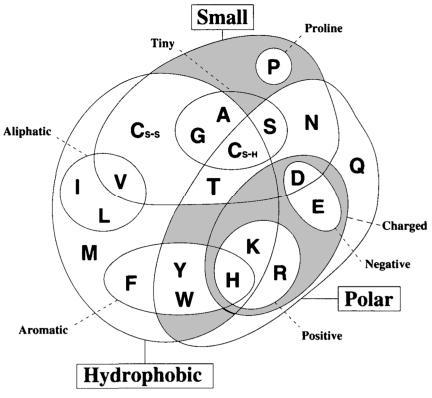
Substitution matrices – substitution biases

Amino acids biochemical properties: nonpolar | polar | basic | acidic

Termination: stop codon

Standard genetic code

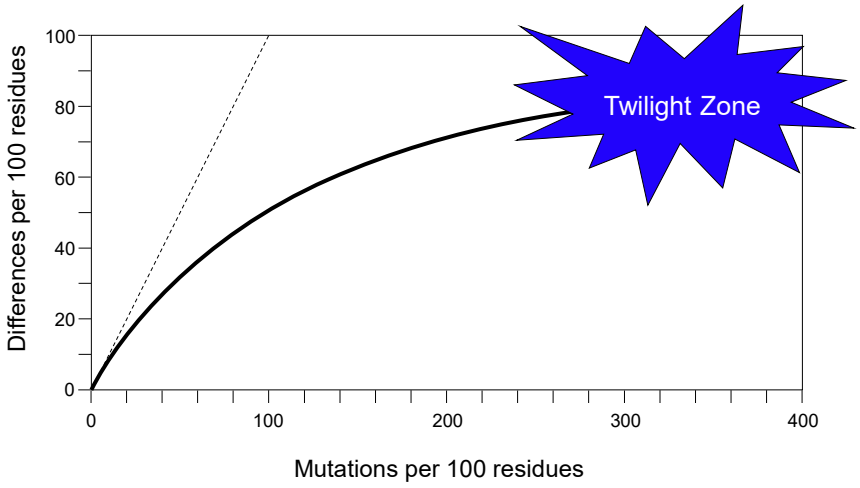
1st base	2nd base T		2nd base C		2nd base A		2nd base G		3rd base
T	TTT	(Phe/F) Phenylalanine	TCT	(Ser/S) Serine	TAT	(Tyr/Y) Tyrosine	TGT	(Cys/C) Cysteine	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) Leucine	TCA		TAA	Stop (*Ochre*)[B]	TGA	Stop (*Opal*)[B]	A
	TTG[A]		TCG		TAG	Stop (*Amber*)[B]	TGG	(Trp/W) Tryptophan	G
C	CTT	(Leu/L) Leucine	CCT	(Pro/P) Proline	CAT	(His/H) Histidine	CGT	(Arg/R) Arginine	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CTG[A]		CCG		CAG		CGG		G
A	ATT	(Ile/I) Isoleucine	ACT	(Thr/T) Threonine	AAT	(Asn/N) Asparagine	AGT	(Ser/S) Serine	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	ATG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GTT	(Val/V) Valine	GCT	(Ala/A) Alanine	GAT	(Asp/D) Aspartic acid	GGT	(Gly/G) Glycine	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GTG		GCG		GAG		GGG		G



From https://en.wikipedia.org/wiki/DNA_codon_table

From Livingstone & Barton, *CABIOS* **9**: 745-56, 1993

Substitution matrices – mutational saturation

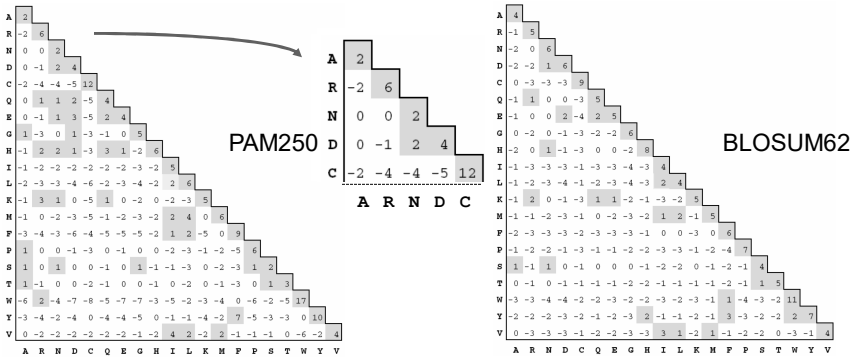


Substitution matrices – theory

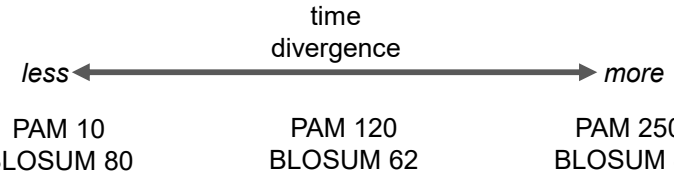
- Need - scoring systems that models sequence change over evolutionary time
- Favour matching identical or related amino acids
- Penalize poorly matched amino acids or gaps
- Take into consideration the relative abundance of amino acids in proteins

Amino acid substitution matrices

- PAM = Point Accepted Mutations (accepted point mutations)
  derived from trusted alignments between closely related sequences
- BLOSUM = Blocks Amino Acid Substitution Matrices
  derived from the BLOCKS database (Pietrokovski et al., 1997; doi: 10.1093/nar/24.1.197)
  ungapped multiple alignments of segments (3 - 60aa in length) of most conserved regions of related proteins.



Amino acid substitution matrices – PAM v. BLOSUM

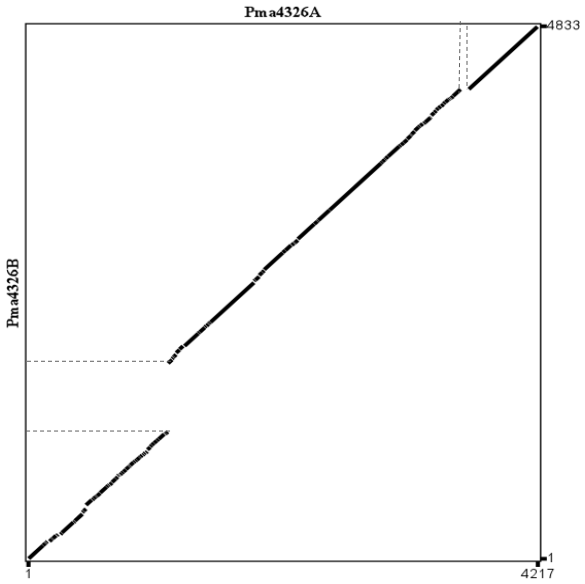


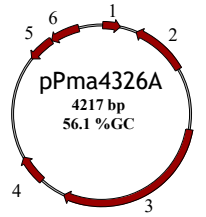
PAM numbers
- evolutionary time
- greater number = greater time since common ancestry for sample

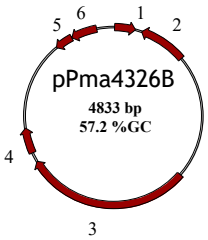
BLOSUM numbers
- sequence similarity
- greater number = greater level of sequence similarity for sample

How to generate an alignment?

EMBOSS Dotmatcher



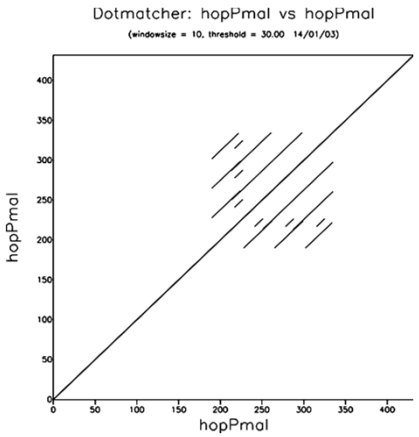


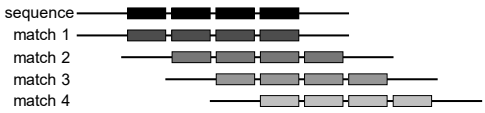


11

Dot Matrix Alignment

EMBOSS Dotmatcher





12

Heuristic methods for aligning sequences

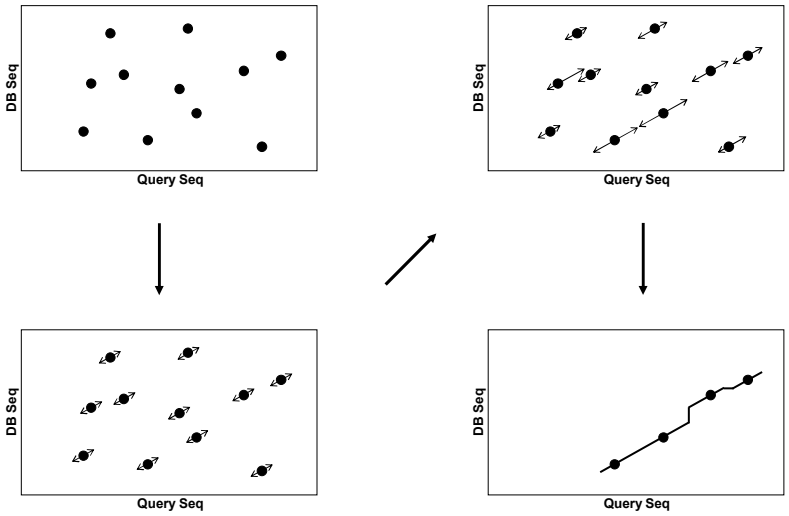
- Heuristic method – exploratory problem solving in which feedback from current result guides future analytical direction.
- Heuristic alignment methods search only a small fraction of the cells in possible search space, while still looking at all the high scoring alignments.
- Heuristic methods are not guaranteed to find the optimal solution, but are much faster (>>50x; necessary if we want to search against the NCBI nr/nt sequence database of more than **388 billion** nucleotides…).
- 2 best known approaches:
  - **BLAST**
  - FASTA

Basic Local Alignment Search Tool (BLAST)

- Program for heuristically finding High Scoring Segment Pairs (HSPs) between a query sequence and a target database.
- Concept
  - true matches very likely contain short stretches of identities
  - short stretches can be seeds for extending the alignment
  - short seed sequences permit preprocessing of queries
- Trade off sensitivity for speed.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215: 215:403

BLAST algorithm – concept



BLAST algorithm – list step

Extract words from query sequence and make expanded list of related words

```
>query
ASGDAAGVSEQTPKLAQYLADKPEHPLNRQRLDAKHEAFKIVVLQALTENCQDAPACNELRELAE
ASG
 SGD
  GDA
   DAA
    AAG
     . . .
```

words

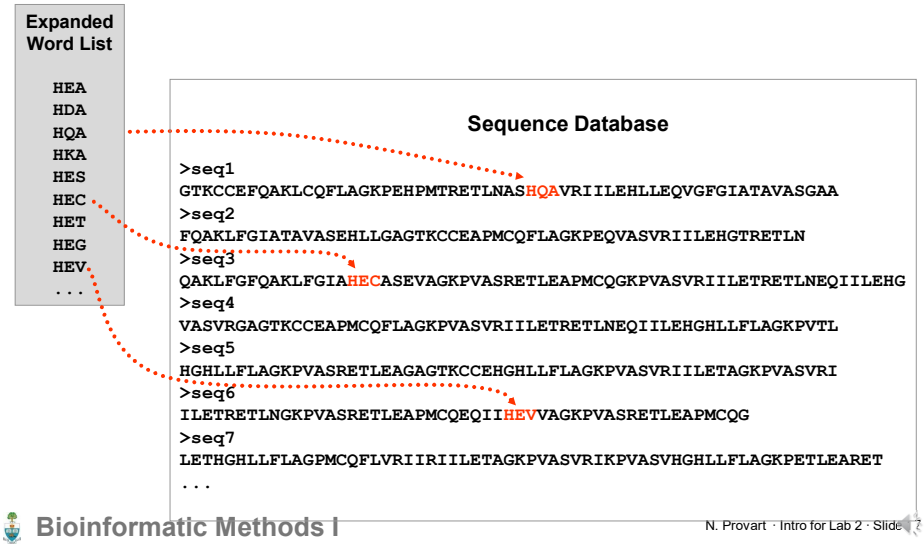
BLOSUM62 matrix

Threshold Word Score = 13

	H	E	A	score
word				
H E A	8	5	4	17
synonyms				
H D A	8	2	4	14
H Q A	8	2	4	14
H K A	8	1	4	13
H E S	8	5	1	14
H E C	8	5	0	13
H E T	8	5	0	13
H E G	8	5	0	13
H E V	8	5	0	13
unacceptable synonym				
Y E A	2	5	4	11

BLAST algorithm – seed step

Scan the selected database for matches to the expanded word list



High-scoring Segment Pair

BLAST algorithm – extend step

```
>query
ASGDAAGVSEQTPKLAQYLADKPEHPLNRQRLDAKHEAFKIVVLQALTENCQDAPACNELRELAE
>seq1
GTKCCEFQAKLCQFLAGKPEHPMTRETLNASHQAVRIILEHLLEQVGFGIATAVASGAA
```

Align at seed word.

```
ASGDAAGVSEQTPKLAQYLADKPEHPLNRQRLDAKHEAFKIVVLQALTENCQDAPACNELRELAE
                                   H+A
    GTKCCEFQAKLCQFLAGKPEHPMTRETLNASHQAVRIILEHLLEQVGFGIATAVASGAA
```

Extend sequence to identify HSP = High-scoring Segment Pair

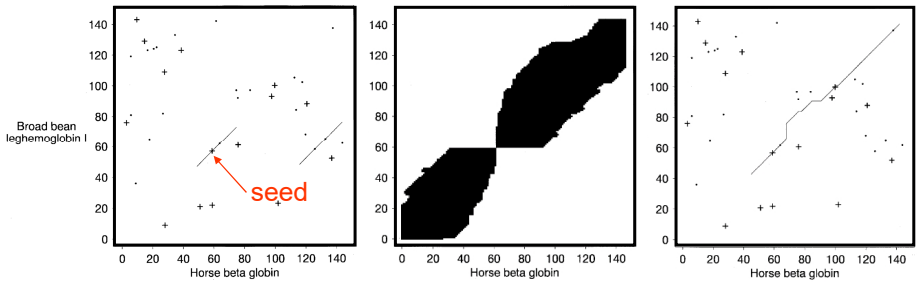
```
ASGDAAGVSEQTPKLAQYLADKPEHPLNRQRLDAKHEAFKIVVLQALTENCQDAPACNELRELAE
            KL Q+LA KPEHP+ R+ L+A H+A +I++   L +
    GTKCCEFQAKLCQFLAGKPEHPMTRETLNASHQAVRIILEHLLEQVGFGIATAVASGAA
```

Stop extension of HSP when quality of the alignment reaches a threshold value. Calculate significance.

```
   KLAQYLADKPEHPLNRQRLDAKHEAFKIVVLQALTE
   KL Q+LA KPEHP+ R+ L+A H+A +I++   L +
   KLCQFLAGKPEHPMTRETLNASHQAVRIILEHLLEQ

Score = 42.0 bits (97), Expect = 0.004
Identities = 17/36 (47%), Positives = 26/36 (72%)
```

BLAST algorithm – gaps



seed

```
Leghemoglobin  43 FSFLKDSAGVVDSPKLGAHAEKVFGMVRDSAVQLRATGEVV--LDGKDGS------  90
                  F  L +    V+ +PK+ AH +KV          L + GE V  LD   G+
Beta globin    45 FGDLSNPGAVMGNPKVKAHGKKV----------LHSFGEGVHHLDNLKGTFAALSE  90

Leghemoglobin  91 IHIQKGVLDP-HFVVVKEALLKTIKEASGDKWSEELSAAWEVAYDGLATAI 140
                    +H  K  +DP +F ++   L+  +    G  ++ EL A+++    G+A A+
Beta globin    91 LHCDKLHVDPENFRLLGNVLVVVLARHFGKDFTPELQASYQKVVAGVANAL 141
```

Altschul, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25: 3389-3402

BLAST - programs

Program	Query	Database	Alignment	# Searches	Uses
blastn	DNA	DNA	DNA	1	find homologous DNA sequences
tblastx	DNA	DNA	protein	36	find homologous proteins from unannotated query and db sequences
blastx	DNA	protein	protein	6	identify proteins in query DNA sequence
tblastn	protein	DNA	protein	6	find homologous proteins in unannotated DNA DB
blastp	protein	protein	protein	1	find homologous proteins

```
                    5'-GTCACGTTACCGGTGGCCGAACAGGCCCGTCATGAAGT-3'
1st reading frame → V  T  L  P  V  A  E  Q  A  R  H  E  V
 2nd reading frame → S  R  Y  R  W  P  N  R  P  V  M  K  X
  3rd reading frame → H  V  T  G  G  R  T  G  P  S  *  S

                    5'-GTCACGTTACCGGTGGCCGAACAGGCCCGTCATGAAGT-3'
                    3'-CAGTGCAATGGCCACCGGCTTGTCCGGGCAGTACTTCA-5'
                       T  V  N  G  T  A  S  C  A  R  *  S  T ← 4th reading frame
                    X  *  T  V  P  P  R  V  P  G  D  H  L ← 5th reading frame
                       D  R  *  R  H  G  F  L  G  T  M  F ← 6th reading frame
```

BLAST – databases

Protein Databases	
nr	Non-redundant GenBank CDS translations + PDB + SwissProt + PIR + PRF
swissprot	Last major release of the SWISS-PROT protein sequence database
pat	Proteins from the Patent division of GenBank.
month	All new or revised GenBank CDS translations + PDB + SwissProt + PIR + PRF released in the last 30 days.
pdb	Sequences derived from the 3-dimensional structure records from the Protein Data Bank

Nucleotide Databases	
nr/nt	All GenBank + EMBL + DDBJ + PDB + RefSeq sequences (but no EST, dbSTS, GSS, WGS,TSA or phase 0, 1 or 2 HTGS sequences).
est	Database of GenBank + EMBL + DDBJ sequences from EST division
refseq_rna	NCBI transcript reference sequences
refseq_representative_genomes	Reference and representative genomes selected from the NCBI Refseq Genomes database
gss	Genome Survey Sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences.
htgs	Unfinished High Throughput Genomic Sequences: phases 0, 1 and 2. Finished, phase 3 HTG sequences are in nr.
pat	Nucleotides from the Patent division of GenBank.
pdb	Sequences derived from the 3-dimensional structure records from Protein Data Bank.
tsa	Transcriptome Shotgun Assembly (TSA) database is an archive of computationally assembled mRNA sequences
sra	Search for sequences associated with a particular SRA (sequence read archive) accession, scientific name, or taxonomic identifier
dbsts	Database of Sequence Tag Site entries from the STS division of GenBank + EMBL + DDBJ.
refseq_genomes	NCBI Refseq genomes across all taxonomy groups. Contains only the top-level sequences, i.e. chromosomal sequences where available (but not the contigs used to assemble them)
wgs	Assemblies of Whole Genome Shotgun sequences

PSI-BLAST – position-specific iterated-BLAST

Motif or profile search methods are often more sensitive than pairwise comparisons at detecting distant relationships.

Most useful for finding protein families.

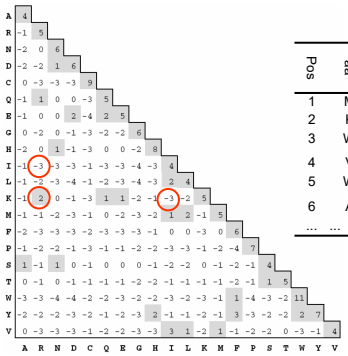
Process

- Create a multiple sequence alignment from BLAST output
- Use the MSA to automatically create a position-specific scoring matrix (PSSM)
  - generated by identifying conserved columns in MSA
- Use PSSM to score BLAST search
- Iterate

PSI-BLAST

```
730496    66 FTVDENGQMSATAKGRVRLFNNWDVCADMIGSFTDTEDPAKFKMKYWGVASFLQKGNDDH 125
200679    63 FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDH 122
206589    34 FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDH  93
2136812    2        MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH  53
132408    65 FKIEDNGKTTATAKGRVRILDKLELCANMVGTFIETNDPAKYRMKYHGALAILERGLDDH 124
267584    44 FSVDESGKVTATAHGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYWGAASYLQTGNDDH 103
267885    44 FSVDGSGKVTATAQGRVIILNNWEMCANMFGTFEDTPDPAKFKERYWGAAAYLQSGNDDH 103
8777608   63 FTIHEDGAMTATAKGRVIILNNWEMCADMMATFETTPDPAKFRMRYWGAASYLQTGNDDH 122
6687453   60 FKVEEDGTMTATAIGRVIILNNWEMCANMFGTFEDTEDPAKFKMKYWGAAAYLQTGYDDH 119
10697027  81 FKVQEDGTMTATATGRVIILNNWEMCANMFGTFEDTEEPARFKMKYWGAAAYLQTGYDDH 140
13645517   1                                MVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH  32
13925316  36 FSVDGSGKMTATAQGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYWGAAAYLQSGNDDH  97
131649    65 YTVEEDGTMTASSKGRVKLFGFWVICADMAAQYTDPTTPAKMYMTYQGLASYLSSGGDNY 126
```

R,I,K C D,E,T K,R,T N,L,Y,G



Pos	aa	Amino Acid																			
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	M	-1	-2	-2	-3	-2	-1	-2	-3	-2	**1**	**2**	-2	**6**	0	-3	-2	-1	-2	-1	**1**
2	K	-1	**1**	0	**1**	-4	**2**	**4**	-2	0	-3	-3	**3**	-2	-4	-1	0	-1	-3	-2	-3
3	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	0	-4	-3	-3	**12**	**2**	-3
4	V	0	-3	-3	-4	-1	-3	-3	-4	-4	**3**	**1**	-3	**1**	-1	-3	-2	0	-3	-1	**4**
5	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	0	-4	-3	-3	**12**	**2**	-3
6	A	**5**	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	**1**	0	-3	-2	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

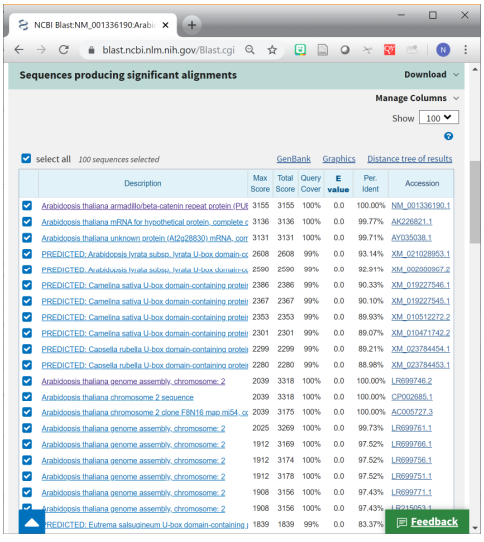
Pevsner 2003 Bioinformatics and Functional Genomics. Wiley-Liss

Evaluation of BLAST results

Is a DB sequence homologous to the query?

- significant expect values
- reciprocal best hit
- similar sizes
- common motifs
- reasonable multiple sequence alignment
- similar 3D structures

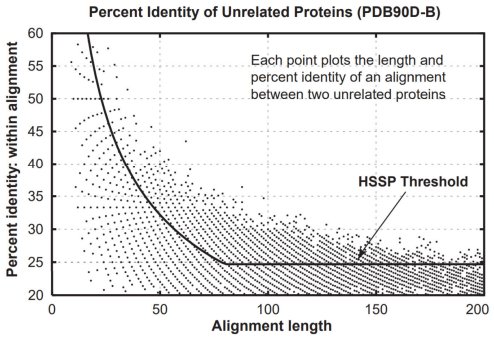
Is one DB hit better than another?



Statistical evaluation – sequence identity?

Why not use sequence identity?

- distribution not well understood
- difficulty with shared domains that do not stretch over length of sequence
- false positive rate
- ignores gaps and conservative vs. radical substitutions



Brenner et al. 1998 PNAS 95:6073

Statistical evaluation – bit score

BLAST reports two bit scores, *S* and *R*

a, b defaults are 1, -2 for Blastn; a slightly different formula and substitution matrices are used for protein bit scores

Raw bit scores (*R*)

$$R = aI + bX - cO - dG$$

I = # identities in the alignment
X = # mismatched residues
O = # gaps
G = # of '-' (length of gap)

a = reward for each identity
b = 'reward' for each mismatch
c = gap opening penalty
d = penalty for each '-'

Can be adjusted manually in Blast

Statistical evaluation – bit score

Normalized bit scores (S)

$$S = (\lambda R - \ln K) / (\ln 2)$$

- λ and K are normalizing parameters
  - λ is a scale factor which converts pairwise match scores to probabilities
  - K is a proportionality constant to correct for the number of sequence comparisons
- Makes bits scores (and E-values) independent of the scoring system

Available from Blast Search Summary

Statistical evaluation – E value

Expect (E) values – best measure of significance

Converts a bit score into a probability

Depend upon

- Bit Score (S)
- Effective length of query (m)
- Effective length of database (n)

$$E = mn2^{-S}$$

Probability of finding a database match as good as or better than your query by chance.

How Good is My Hit?

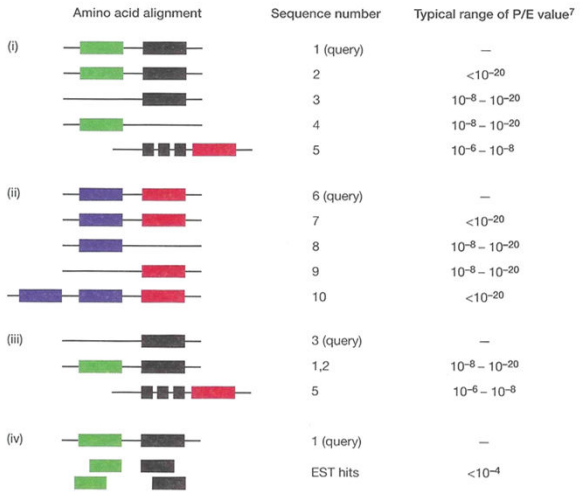
Use identity? No!

Use bit score: better

Use E value: best

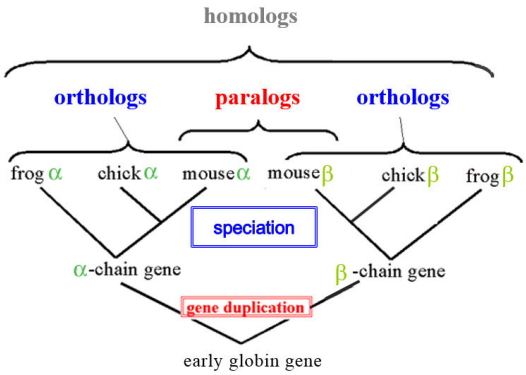
- The E value is a probability value that is based on the number of different alignments with scores at least as good as that observed, which are expected to occur simply by chance.
- The lower the E value, the more significant the score. This is by far the best metric to use since results of different searches in the same database can be readily compared.
- Note that E value is dependent on the size of the database (n) and the length of the query sequence (m). *The same sequence searched on different databases containing identical hit sequences would result in different E values being reported for those sequences.*

What is a good E value?



From Mount, Bioinformatics (2001), p. 497

Orthology and Paralogy



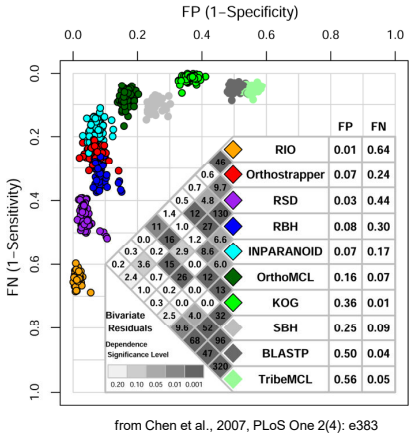
- Orthology can be used to identify conserved residues within genes and proteins
- In addition, comparative genomic methods can be applied to intron sequences and promoters to identify parts of these that are conserved and hence potentially functionally important

from http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Orthology.html

Methods for determining orthology in genomic sequences

- TBLASTX or BLASTP – take reference genome and blast against other genomes, and take region (gene) with best e-value (above a threshold) as orthologous region. Problem: what if blasting in other direction identifies a match in reference genome that is better? Which is the ortholog?
- Reciprocal Best Hit (RBH) method – addresses the above issue but can get confounded by rampant domain swapping that has occurred, esp. in eukaryotic genomes → lots of false negatives
- Phlyogenetic-based methods such as RIO, Orthostrapper and RSD
- BLASTP-based methods, such as InParanoid, OrthoMCL, KOG: these use BLASTP followed by Markov or other Clustering methods

Overview of the methods: which is best?



from Chen et al., 2007, PLoS One 2(4): e383

Other tools are available, e.g. OrthoFinder2
(Emms & Kelly, 2019, https://dx.doi.org/10.1186/s13059-019-1832-y)

→ what are FP and FN rates for any tool you might want to use?

Ortholog databases

Clusters of Orthologous Groups (COG) and euKaryotic Orthologous (KOG) Groups: http://www.ncbi.nlm.nih.gov/COG/ *several species, older

HieranoiDB: http://hieranoidb.sbc.su.se/ *66 species, slightly older
Kaduk M, Riegler C, Lemp O, Sonnhammer EL. HieranoiDB: a database of orthologs inferred by Hieranoid. Nucleic Acids Res. 2017, 45(Database issue), D687-D690. PMID: 27742821.

OrthoMCL DB: http://www.orthomcl.org/ *many species, slightly out-of-date
Feng Chen, Aaron J. Mackey, Christian J. Stoeckert, Jr and David S. Roos. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Research 2006 34(Database Issue):D363-D368

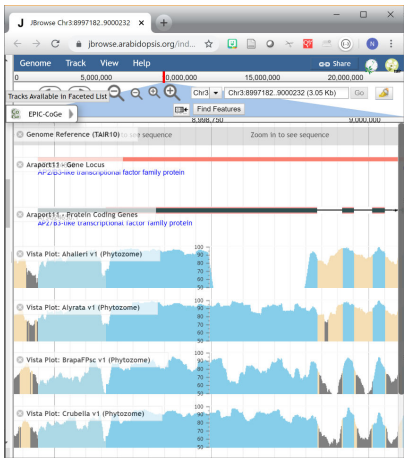
InParanoid DB: http://inparanoid.sbc.su.se/cgi-bin/index.cgi *273 species, from 2013
Remm M, Storm CEV and Sonnhammer ELL (2001). Automatic Clustering of Orthologs and In-paralogs from Pairwise Species Comparisons. JMB, 314:1041-1052.

CoGe: http://genomeevolution.org/ 50,000+ genomes, up-to-date; syntenty tools!
Lyons E ~ Lisch D (2008) Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar and grape: CoGe with rosids, Plant Phys 148, pp. 1772–1781.

You may find others → how up-to-date are these, genome versions, etc.?

Tools for comparative genomics & genome browsing

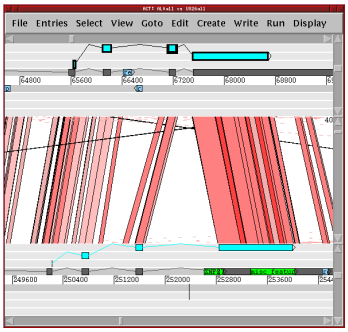
GBrowse/JBrowse is standard for many model organisms



ACT (Artemis Comparison Tool) standalone tool that allows cross-genome comparisons

http://www.sanger.ac.uk/science/tools/artemis

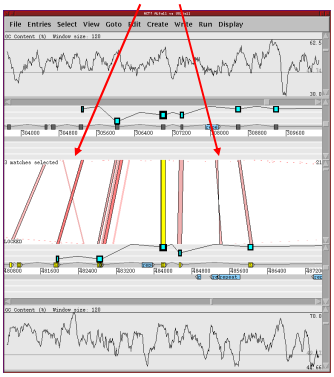
allows rearrangements and syntenic blocks to be easily visualized



Genome comparisons and synteny

- Synteny is the preservation of gene order on chromosomes of related species
- During evolution, genomic rearrangements can separate two loci → result is a loss of synteny between them
- Translocations can also join two previously separate pieces of chromosomes (rare event) → results in a gain of synteny between loci
- Synteny can be useful in the case of many-to-many or one-to-many ortholog mappings, for determining the "true" ortholog, and also identifying translocations/ inversions – these show up as blocks which cross other blocks, and as "X" shaped figures in e.g. ACT

Potential translocations/inversions



Overview

Substitution Matrices

Alignment Methods

- **Dot Matrix**
- Dynamic Programming
  - Global Alignment
  - Local Alignment
- Hidden Markov Models
- Heuristic Alignment - k-tuple
  - **BLAST**
  - FASTA

Evaluation of Significance

Comparative Genomics