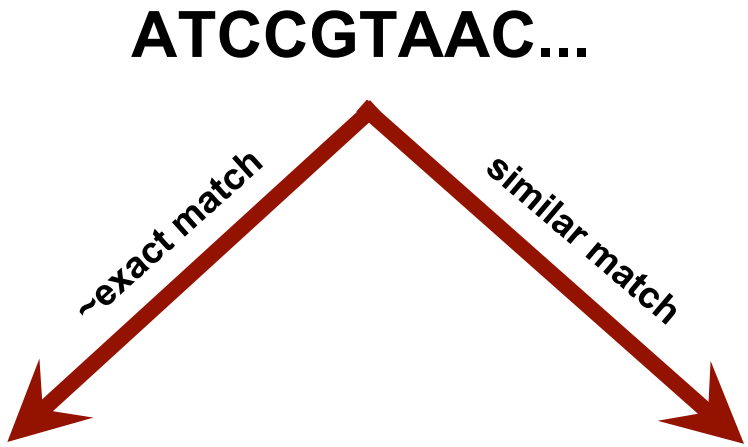
Introduction to Sequence Analysis

ATS 2008
Toronto, Ontario
May 16, 2008

Courtney Harper
Resource for Biocomputing, Visualization, and Informatics
charper@gmail.com

› Sequence analysis provides access to genomics/related information

– because it represents the primary data, accessing genomics Web sites via sequence comparison bypasses problems associated with searching using key words, gene names, various types of accession #s



Access to many types of info about a gene/protein

- browsers
- organism DBs
- specialty DBs (ex: IGTC, GNF expression data)
- proteomics info

Near & distant homologs in multiple species

- primary sequence DBs
- precalculated family DBs
- gene families within a species of interest (ex: GPCRs)

Exact (or nearly exact) matches: finding your gene

Example: searching for a Ldlr knockout

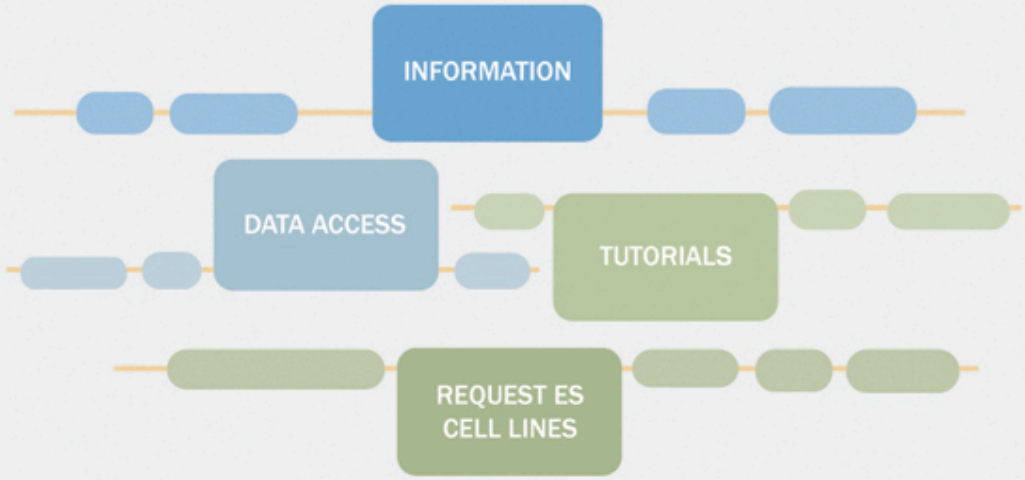
IGTC International Gene Trap Consortium

INFORMATION | DATA ACCESS | TUTORIALS | REQUEST ES CELL LINES

About IGTC

Gene trapping is a high-throughput approach that is used to introduce insertional mutations across the genome in mouse embryonic stem (ES) cells. In addition to generating standard loss-of-function alleles, newer gene trap vectors offer a variety of post-insertional modification strategies for the generation of other experimental alleles.

The International Gene Trap Consortium (IGTC) represents all publicly available gene trap cell lines, which are available on a non-collaborative basis for nominal handling fees. Researchers can search and browse the IGTC database for cell lines of interest using accession numbers or IDs, keywords, sequence data, tissue expression profiles and biological pathways.



Statistics and News

› Statistics: Mar 28, 2008

IGTC cell lines in database:128076 cell lines
Pipeline Status:128075 (100.00%) processed cell lines
Last dbGSS deposit:2008-03-26
Ensembl version:48, NCBI m37, IGTC cell line coverage 8801 (32.04%) genes
Entrez version:NCBI m37, IGTC cell line coverage 11204 (17.53%) genes

› IGTC Switches to Unmasked Genome

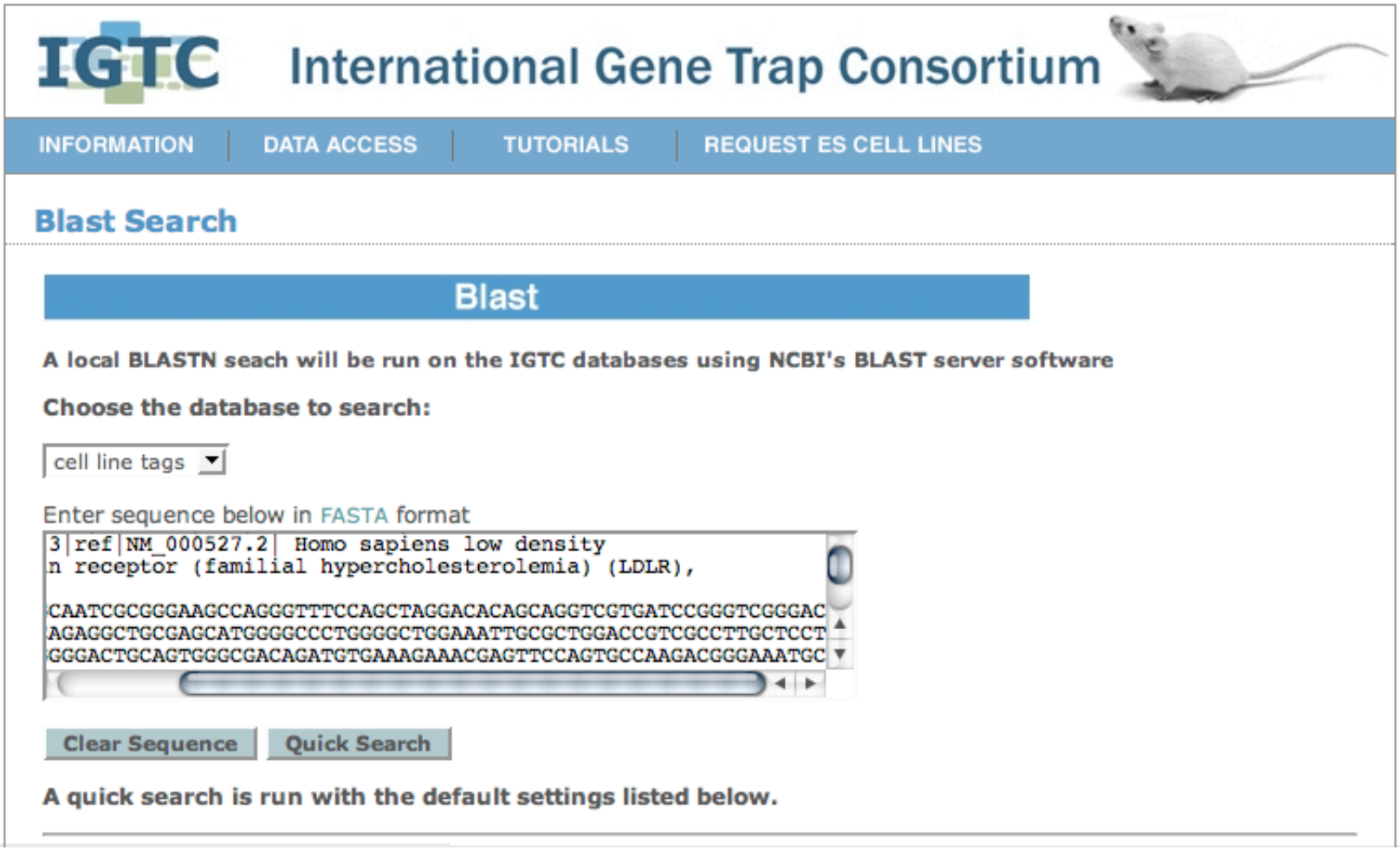
Mar 14, 2008

Today the IGTC began using the unmasked version of the mouse genome for all localization. Prior to today, the IGTC used the masked genome. Overall we expect this change to result in more cell line identifications and therefore more gene associations as sequence tags are reprocessed over the coming weeks. See the pipeline description for more information about cell line processing. The UCSC website also describes how the two genome versions differ.

› *In Situ* Images Available

Example: searching for a Ldlr knockout

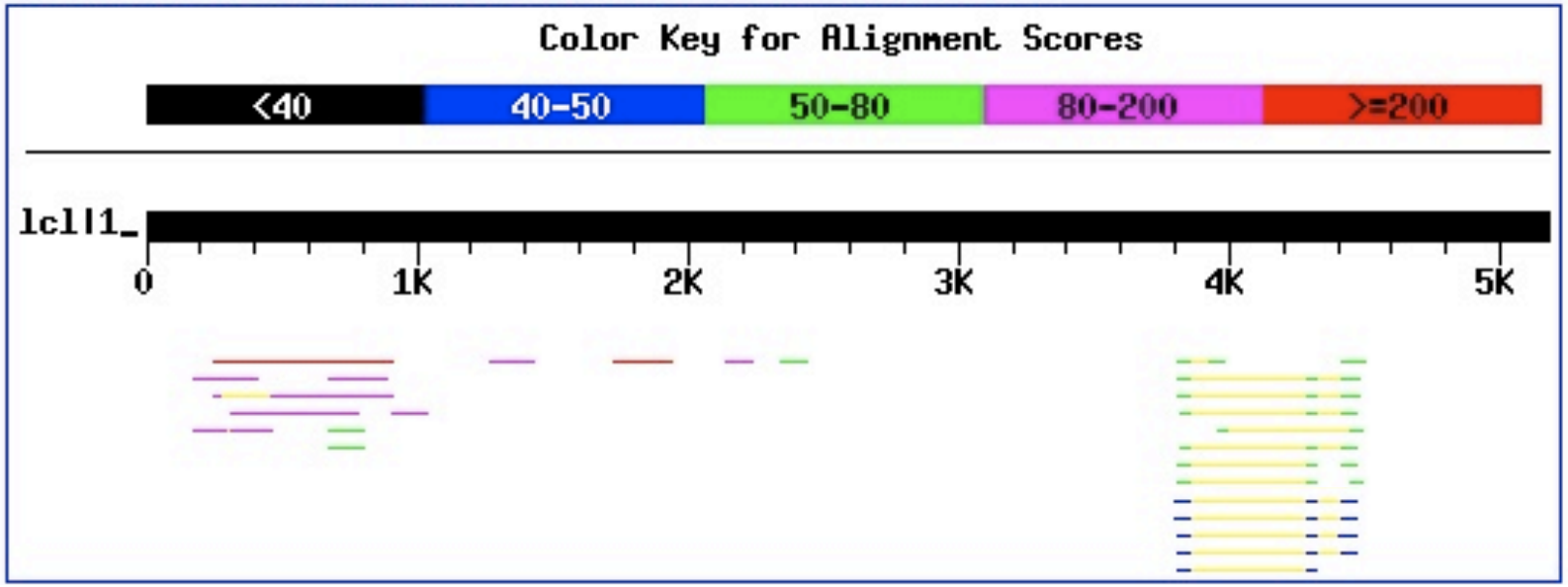
Example: searching for a Ldlr knockout



Example: searching for a Ldlr knockout

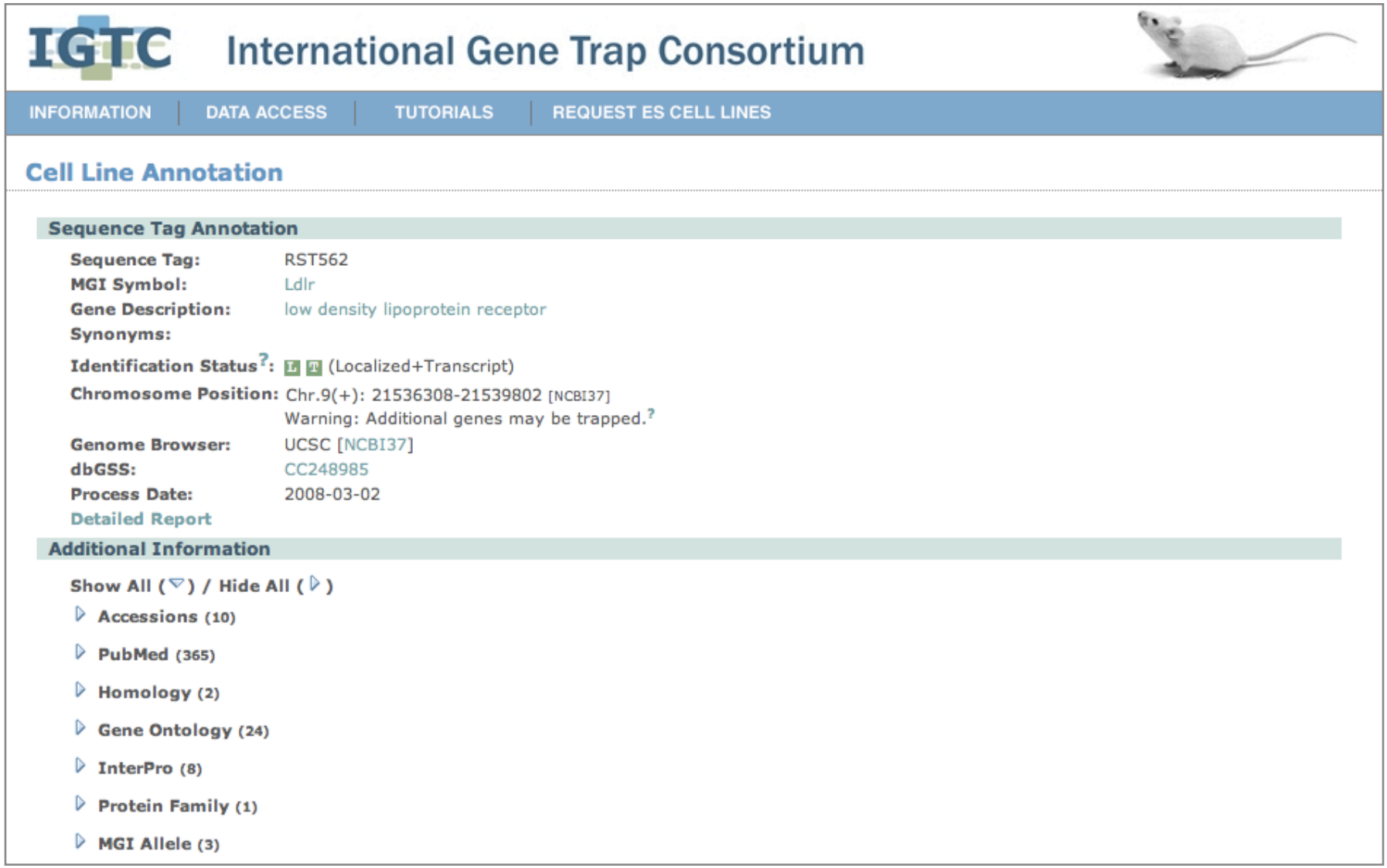
Distribution of 114 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments



Sequences producing significant alignments:		Score (bits)	E Value
RST562	(BG)	354	1e-96
RST050	(BG)	250	2e-65
RST527	(BG)	198	1e-49
RST575	(BG)	185	1e-45
LST106	(BG)	160	4e-38

Example: searching for a Ldlr knockout



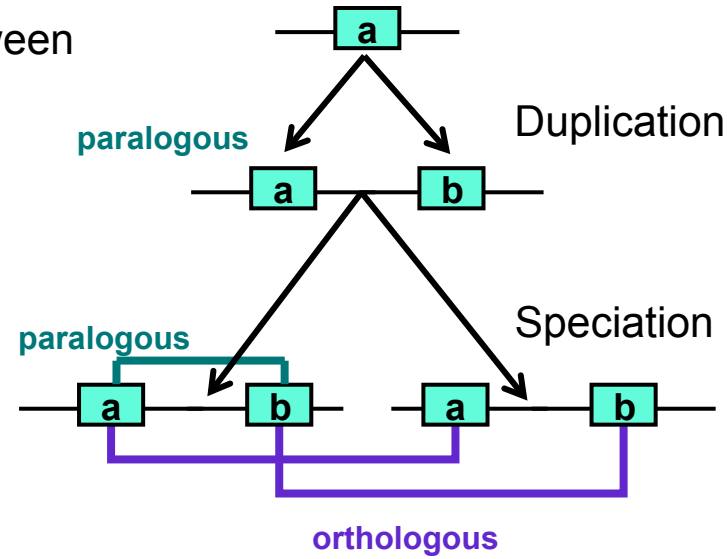
Distant matches:
inferring functional properties

Homology: Sharing of a common ancestor by descent or recombination;

- 2 proteins can't be "significantly" homologous, 60% homologous…
- homology is a conjecture, not an experimental fact and requires knowledge of the evolutionary relationships among the sequences being compared only degrees of similarity can be quantified and explicitly defined

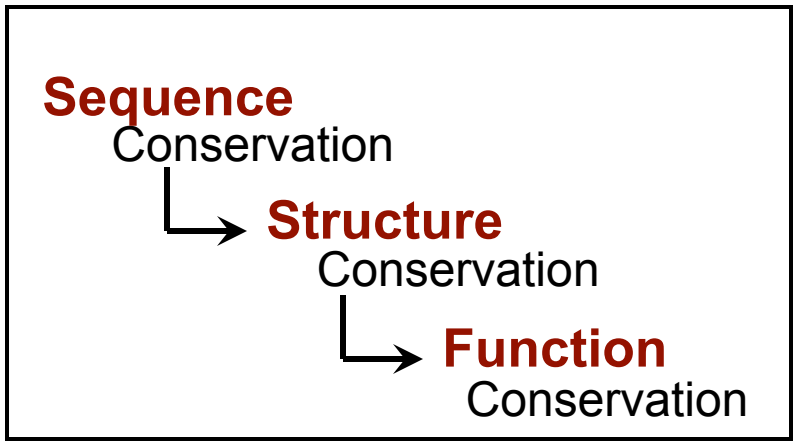
Orthology: sequence similarity as a consequence of a speciation event

Parology: sequence similarity between the descendants of a duplicated ancestral gene



Matches to similar sequences

Involves accessing evolutionary signals



Rule: What is conserved in a gene [protein] family is functionally important

- due to purifying selection driven by functional constraints observable in a background described by the theory of neutral evolution
  - fast enough that pseudogenes rapidly deteriorate over evolutionary timescales
  - in any prokaryotic genome, homologs from more than one distantly related species are detectable for 70-80% of proteins

Application: Comparison of sequence/structures can identify homologous relationships, allowing inference of function based on that relationship

Practical Issues in Sequence Analysis

1. Database searching
   - principles for pairwise comparisons
   - primary tools
2. Using BLAST & interpreting the results
   - BLAST output
   - statistics & interpreting E value
3. Using Psi-BLAST to find distantly related sequences
4. Evaluation using multiple alignments
5. Adding 3D structure information
6. Resources, tools, references
   - posted at the end of this lecture, which is available at http://www.genetrap.org/tutorials/seqanalysis.html

Database Searching

How to find a relationship between a query sequence & sequences in a database?

- correspondence between 2 or more aligned sequences
  - similarity score
  - graphical representations: alignments, motifs, patterns
- candidate homologs evaluated using statistics

Formalizing the problem - pairwise comparisons

Given: two sequences that you want to align

Goal: find the best alignment that can be obtained by sliding one sequence along the other

Requirements:

- a scheme for evaluating matches/mis-matches between any two characters
- a score for insertions/deletions
- a method for optimization of the total score
- a method for evaluating the statistical significance of the alignment

BLOSUM62 matrix

```
...DGAERSWIW...
   ||  |||  ||
...DGWERSDIW...
```

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	**9**	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	**4**	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	**4**	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	**7**	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	**4**	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	**6**	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	**6**	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	**6**	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	**5**	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	**5**	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	**8**	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	**5**	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	**5**	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	**5**	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	**4**	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	**4**	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	**4**	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	**6**	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	**7**	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	**11**

A = Alanine (small and neutral)
K = Lysine (positively charged)
W = Tryptophan (large and hydrophobic)

Scoring systems

The degree of match between two letters can be represented in a matrix and changing the matrix can change the alignment

- Simplest: Identity (unitary) matrix
- Better: Definitions of similarity based on inferences about chemical or biological properties
  - Examples: PAM, Blosum, Gonnet, profiles, HMMs, neural network models, etc.
- Scores often reported w statistics of the form $p_{ab}/q_a\,q_b$
  where p_{ab} is the probability that residue *a* is substituted by residue *b,* and q_a and q_b are the background probabilities for residue *a* and *b* respectively
- Scores historically derived empirically from an evolutionary model describing expected evolutionary change by point mutations (scoring gaps not accommodated in the models)
  - models used to define expected numbers & types of mutations based on evolutionary distance

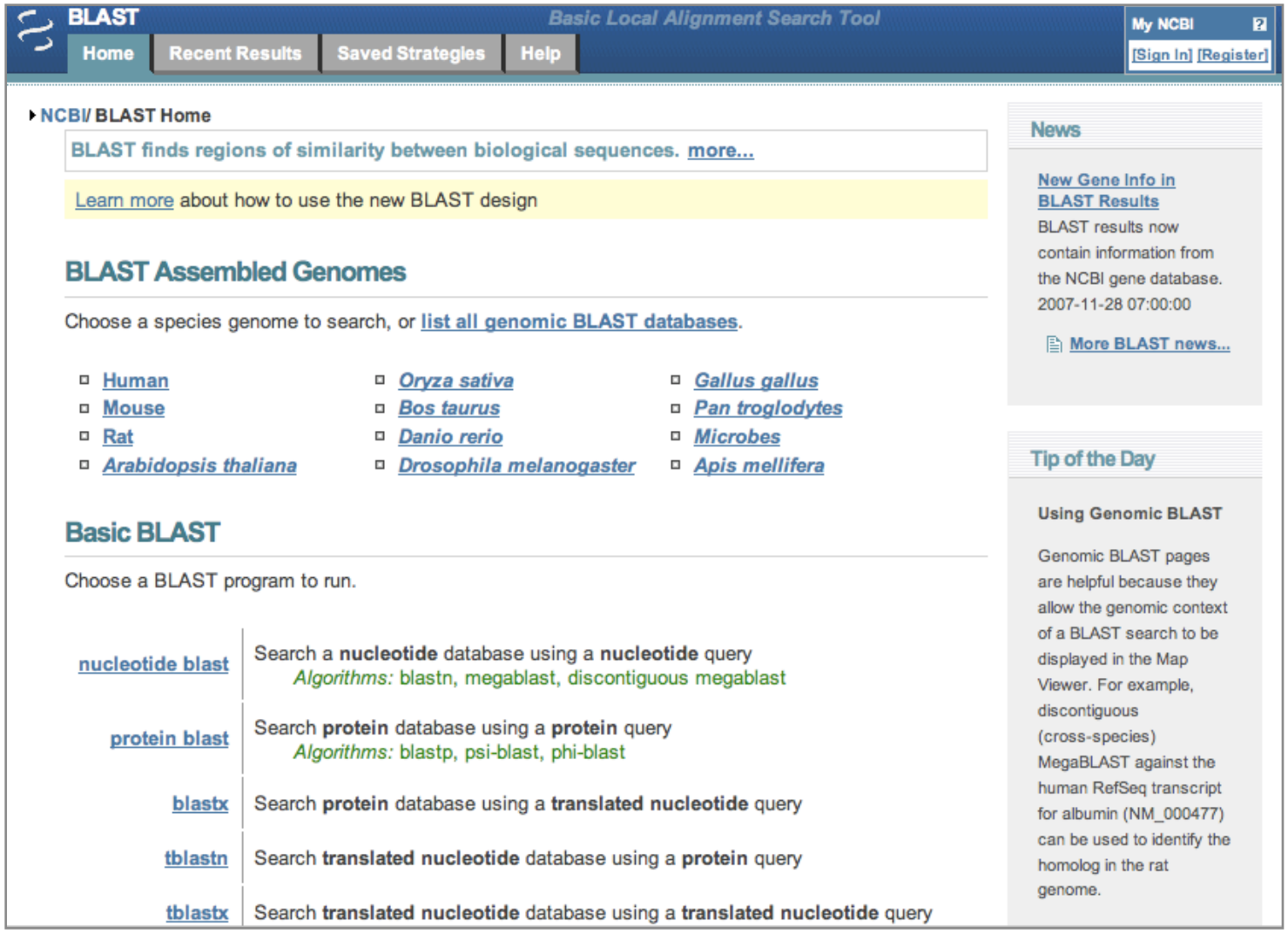
Database searching

Current algorithms handle large DBs quickly, even for divergent relationships

- generates alignments & estimates of statistical significance
- different scoring matrices/other parameters provide tuning
- 2 major heuristic algorithms compromise speed & sensitivity (only slightly compared to more rigorous but slower algorithms such as Smith-Waterman)
  - BLAST generates pairwise alignments, based on rigorous statistical formalism for the significance of these alignments (http://www.ncbi.nlm.nih.gov/BLAST/; http://blast.wustl.edu/)
  - FASTA (http://fasta.bioch.virginia.edu/)
- very fast variants developed for localizing sequences to the genomes (faster because they only look for ~exact matches)
  - BLAT (UC Santa Cruz browser & GNF expression datasets)
  - SSAHA2 (Ensembl)
  - MegaBlast (NCBI genome browser)
  - GMAP (Genentech)

BLAST

(excellent documentation at http://www.ncbi.nlm.nih.gov/BLAST/)



BLAST flavors

blastp compares an amino acid query sequence against a protein sequence database

blastn compares a nucleotide query sequence against a nucleotide sequence database

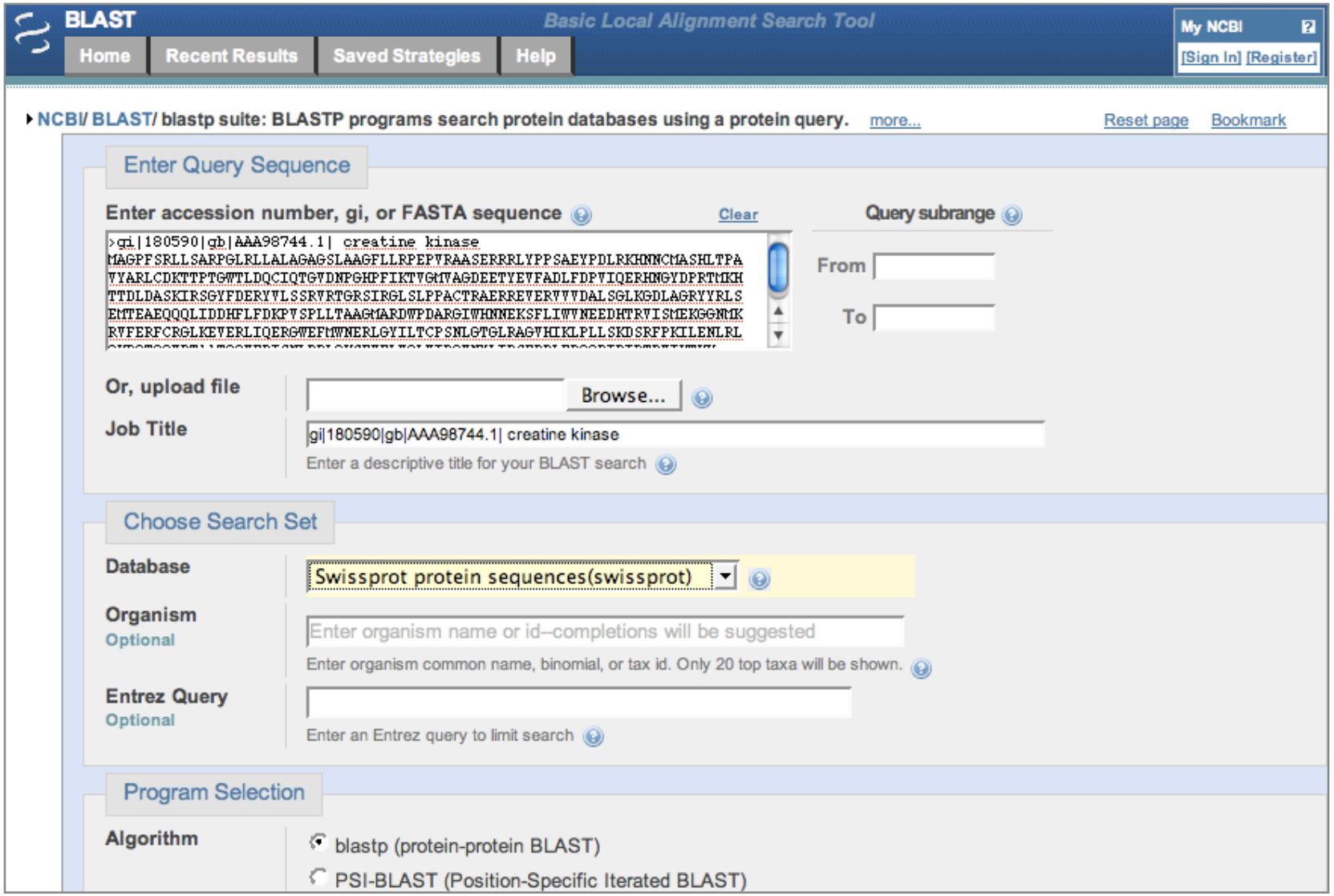
blastx compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database

tblastn compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands)

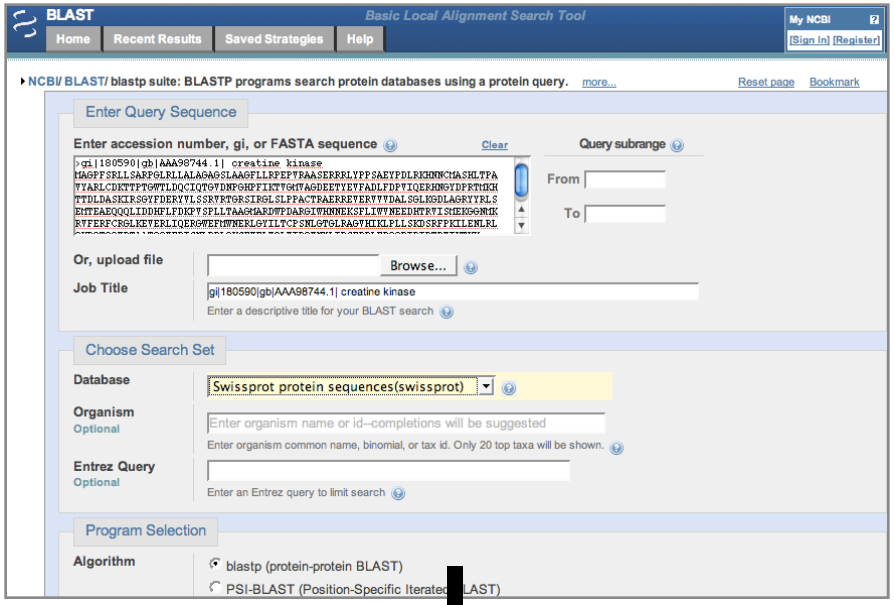
tblastx compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database

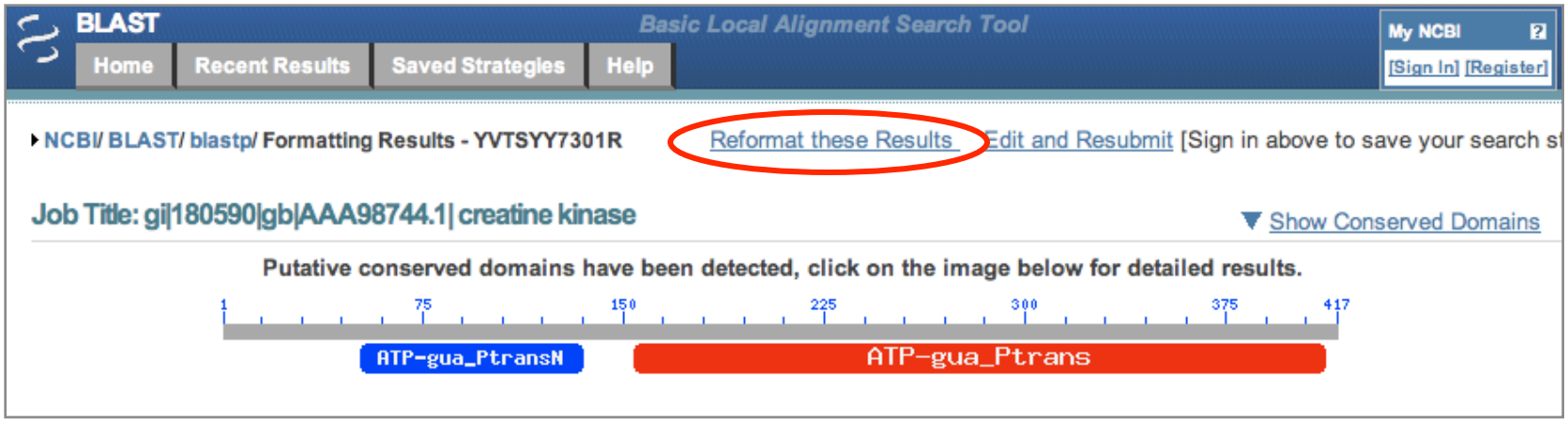
Running BLAST
&
Interpreting Results

Example: highly conserved gene families – creatine kinase

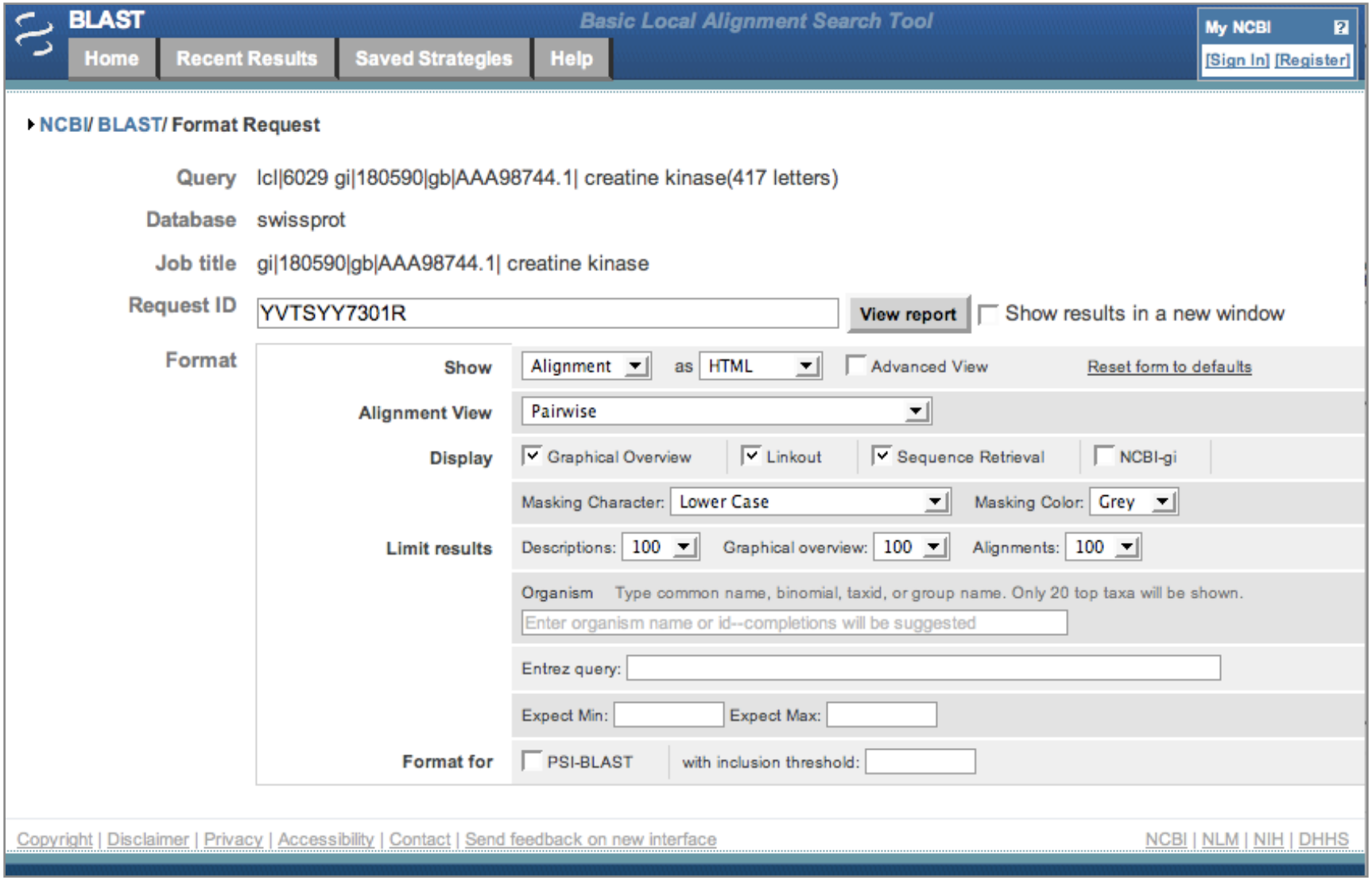


Example: highly conserved gene families – creatine kinase



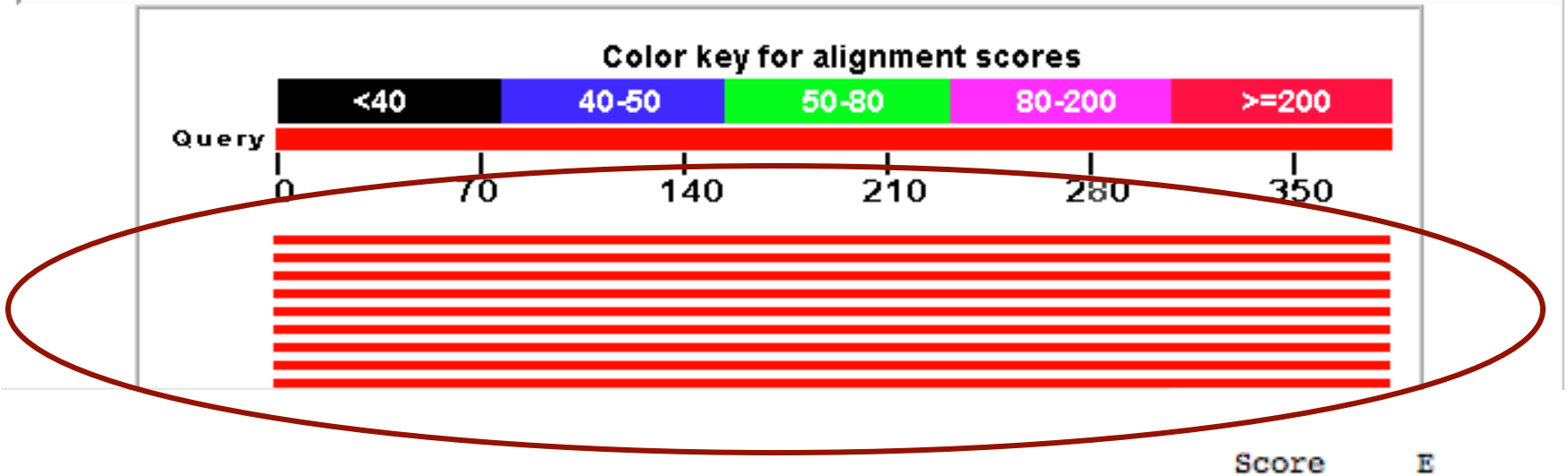


Example: highly conserved gene families – creatine kinase



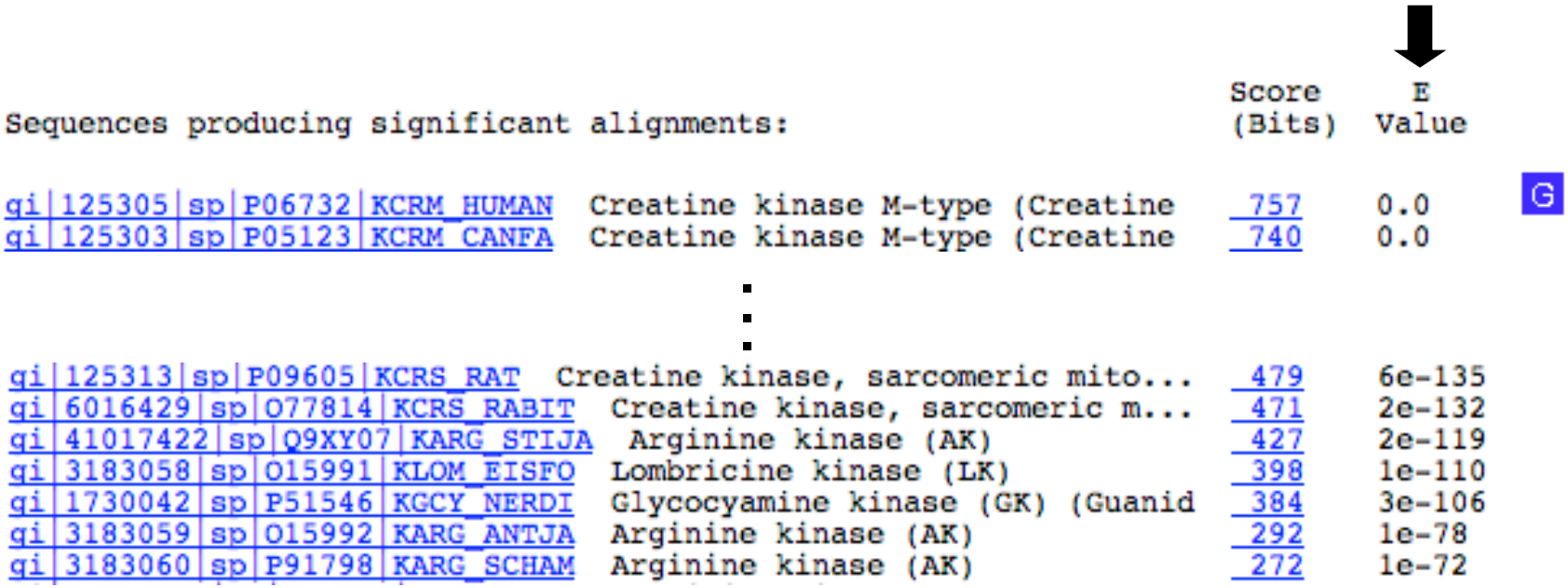
Distribution of 90 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments



Sequences producing significant alignments:		Score (Bits)	E Value
qi|125305|sp|P06732|KCRM_HUMAN	Creatine kinase M-type (Creatine	757	0.0
qi|125303|sp|P05123|KCRM_CANFA	Creatine kinase M-type (Creatine	740	0.0
⋮			
qi|125313|sp|P09605|KCRS_RAT	Creatine kinase, sarcomeric mito...	479	6e-135
qi|6016429|sp|O77814|KCRS_RABIT	Creatine kinase, sarcomeric m...	471	2e-132
qi|41017422|sp|Q9XY07|KARG_STIJA	Arginine kinase (AK)	427	2e-119
qi|3183058|sp|O15991|KLOM_EISFO	Lombricine kinase (LK)	398	1e-110
qi|1730042|sp|P51546|KGCY_NERDI	Glycocyamine kinase (GK) (Guanid	384	3e-106
qi|3183059|sp|O15992|KARG_ANTJA	Arginine kinase (AK)	292	1e-78
qi|3183060|sp|P91798|KARG_SCHAM	Arginine kinase (AK)	272	1e-72

What's an E value?

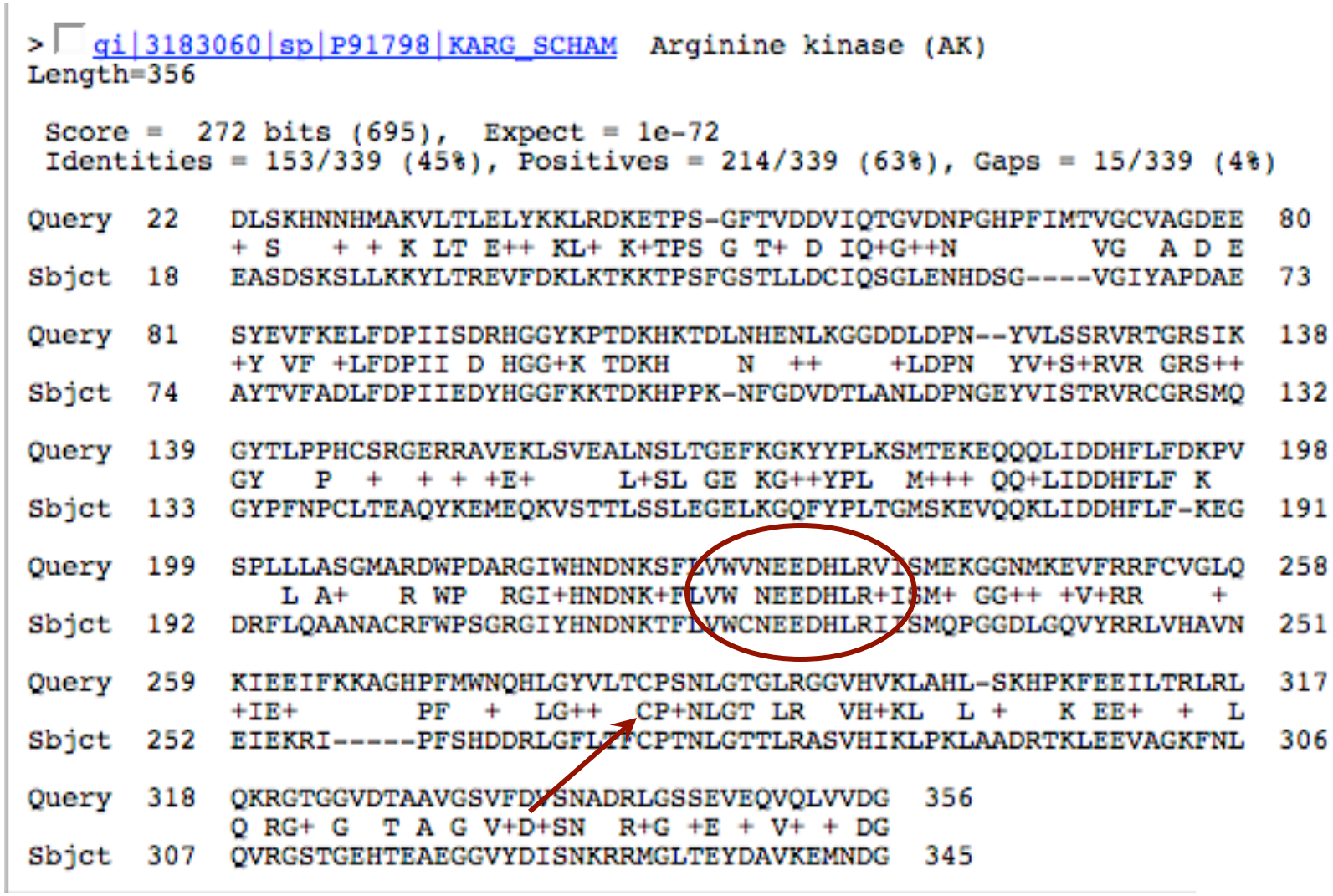


Sequences producing significant alignments:		Score (Bits)	E Value
qi|125305|sp|P06732|KCRM_HUMAN	Creatine kinase M-type (Creatine	757	0.0
qi|125303|sp|P05123|KCRM_CANFA	Creatine kinase M-type (Creatine	740	0.0
⋮			
qi|125313|sp|P09605|KCRS_RAT	Creatine kinase, sarcomeric mito...	479	6e-135
qi|6016429|sp|O77814|KCRS_RABIT	Creatine kinase, sarcomeric m...	471	2e-132
qi|41017422|sp|Q9XY07|KARG_STIJA	Arginine kinase (AK)	427	2e-119
qi|3183058|sp|O15991|KLOM_EISFO	Lombricine kinase (LK)	398	1e-110
qi|1730042|sp|P51546|KGCY_NERDI	Glycocyamine kinase (GK) (Guanid	384	3e-106
qi|3183059|sp|O15992|KARG_ANTJA	Arginine kinase (AK)	292	1e-78
qi|3183060|sp|P91798|KARG_SCHAM	Arginine kinase (AK)	272	1e-72

E value = Expectation value, the number of alignments with scores ≥ than the associated score expected to occur in a database search by chance

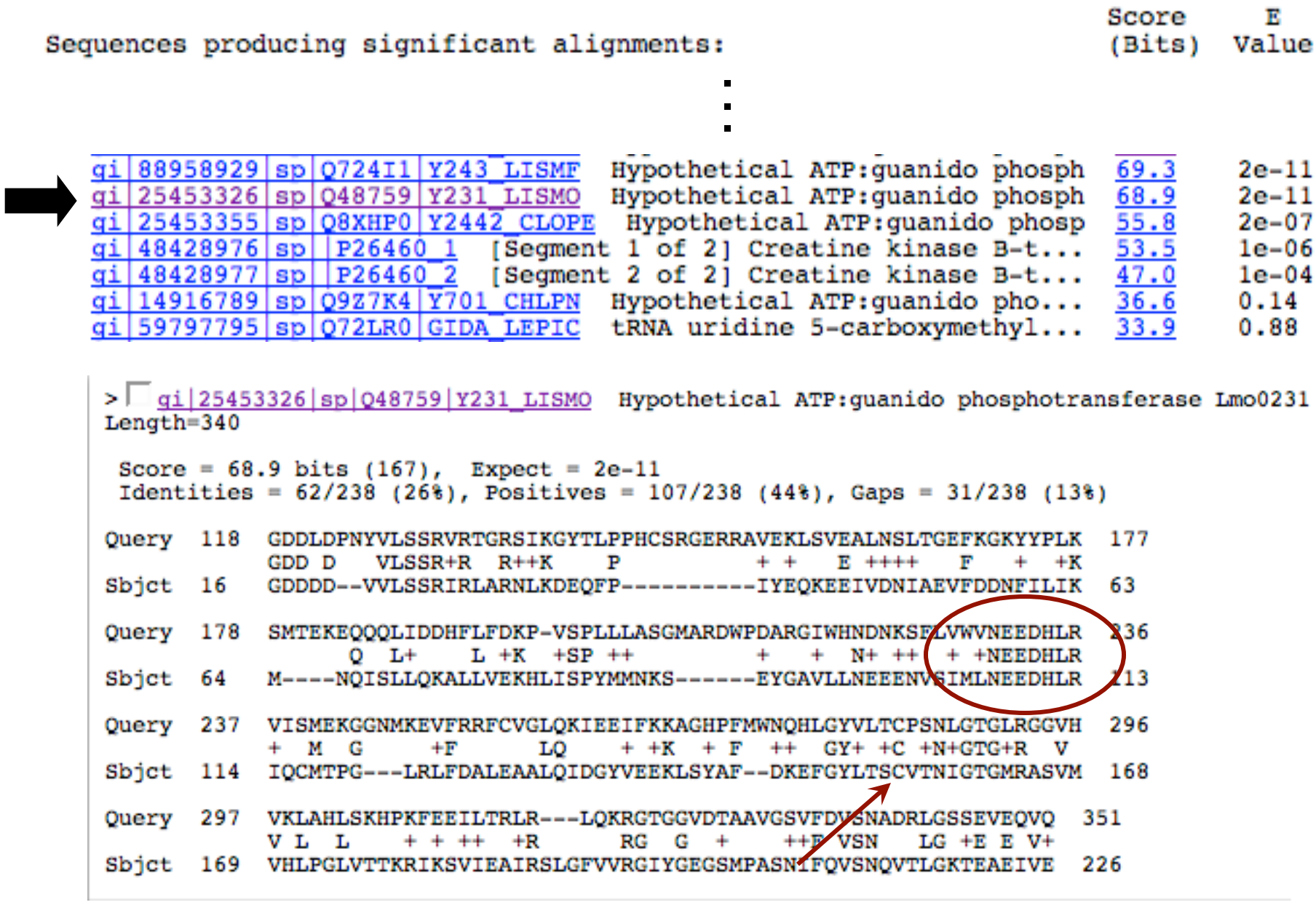
- for E-values, the smaller the better!
- depends on both the size of the alignments & the size of the sequence database
  - E-value INCREASES as the database gets bigger
  - E-value DECREASES as the alignment gets longer

Example: highly conserved gene families – creatine kinase



```
>  qi|3183060|sp|P91798|KARG_SCHAM  Arginine kinase (AK)
Length=356

 Score =  272 bits (695),  Expect = 1e-72
 Identities = 153/339 (45%), Positives = 214/339 (63%), Gaps = 15/339 (4%)

Query  22   DLSKHNNHMAKVLTLELYKKLRDKETPS-GFTVDDVIQTGVDNPGHPFIMTVGCVAGDEE  80
            + S   + + K LT E++ KL+ K+TPS G T+ D IQ+G++N        VG  A D E
Sbjct  18   EASDSKSLLKKYLTREVFDKLKTKKTPSFGSTLLDCIQSGLENHDSG----VGIYAPDAE  73

Query  81   SYEVFKELFDPIISDRHGGYKPTDKHKTDLNHENLKGGDDLDPN--YVLSSRVRTGRSIK  138
            +Y VF +LFDPII D HGG+K TDKH    N  ++    +LDPN  YV+S+RVR GRS++
Sbjct  74   AYTVFADLFDPIIEDYHGGFKKTDKHPPK-NFGDVDTLANLDPNGEYVISTRVRCGRSMQ  132

Query  139  GYTLPPHCSRGERRAVEKLSVEALNSLTGEFKGKYYPLKSMTEKEQQQLIDDHFLFDKPV  198
            GY   P  +  + + +E+     L+SL GE KG++YPL  M+++ QQ+LIDDHFLF K
Sbjct  133  GYPFNPCLTEAQYKEMEQKVSTTLSSLEGELKGQFYPLTGMSKEVQQKLIDDHFLF-KEG  191

Query  199  SPLLLASGMARDWPDARGIWHNDNKSFLVWVNEEDHLRVISMEKGGNMKEVFRRFCVGLQ  258
               L A+   R WP  RGI+HNDNK+FLVW NEEDHLR+I+M+ GG++ +V+RR    +
Sbjct  192  DRFLQAANACRFWPSGRGIYHNDNKTFLVWCNEEDHLRIISMQPGGDLGQVYRRLVHAVN  251

Query  259  KIEEIFKKAGHPFMWNQHLGYVLTCPSNLGTGLRGGVHVKLAHL-SKHPKFEEILTRLRL  317
            +IE+       PF  +  LG++  CP+NLGT LR  VH+KL  L +   K EE+  +  L
Sbjct  252  EIEKRI-----PFSHDDRLGFLTFCPTNLGTTLRASVHIKLPKLAADRTKLEEVAGKFNL  306

Query  318  QKRGTGGVDTAAVGSVFDVSNADRLGSSEVEQVQLVVDG  356
            Q RG+ G  T A G V+D+SN  R+G +E + V+ + DG
Sbjct  307  QVRGSTGEHTEAEGGVYDISNKRRMGLTEYDAVKEMNDG  345
```

Example: highly conserved gene families – creatine kinase



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

gi|88958929|sp|Q724I1|Y243_LISMF  Hypothetical ATP:guanido phosph   69.3    2e-11
gi|25453326|sp|Q48759|Y231_LISMO  Hypothetical ATP:guanido phosph   68.9    2e-11
gi|25453355|sp|Q8XHP0|Y2442_CLOPE  Hypothetical ATP:guanido phosp   55.8    2e-07
gi|48428976|sp|P26460_1  [Segment 1 of 2] Creatine kinase B-t...   53.5    1e-06
gi|48428977|sp|P26460_2  [Segment 2 of 2] Creatine kinase B-t...   47.0    1e-04
gi|14916789|sp|Q9Z7K4|Y701_CHLPN  Hypothetical ATP:guanido pho...   36.6    0.14
gi|59797795|sp|Q72LR0|GIDA_LEPIC  tRNA uridine 5-carboxymethyl...   33.9    0.88

>gi|25453326|sp|Q48759|Y231_LISMO  Hypothetical ATP:guanido phosphotransferase Lmo0231
Length=340

 Score = 68.9 bits (167),  Expect = 2e-11
 Identities = 62/238 (26%), Positives = 107/238 (44%), Gaps = 31/238 (13%)

Query  118  GDDLDPNYVLSSRVRTGRSIKGYTLPPHCSRGERRAVEKLSVEALNSLTGEFKGKYYPLK  177
            GDD D   VLSSR+R  R++K    P          + +   E ++++   F   +  +K
Sbjct  16   GDDDD--VVLSSRIRLARNLKDEQFP----------IYEQKEEIVDNIAEVFDDNFILIK  63

Query  178  SMTEKEQQQLIDDHFLFDKP-VSPLLLASGMARDWPDARGIWHNDNKSFLVWVNEEDHLR  236
                  Q  L+    L +K  +SP ++          +   +  N+ ++  + +NEEDHLR
Sbjct  64   M----NQISLLQKALLVEKHLISPYMMNKS------EYGAVLLNEEENVSIMLNEEDHLR  113

Query  237  VISMEKGGNMKEVFRRFCVGLQKIEEIFKKAGHPFMWNQHLGYVLTCPSNLGTGLRGGVH  296
            +  M  G     +F      LQ    + +K  + F  ++  GY+ +C +N+GTG+R  V
Sbjct  114  IQCMTPG---LRLFDALEAALQIDGYVEEKLSYAF--DKEFGYLTSCVTNIGTGMRASVM  168

Query  297  VKLAHLSKHPKFEEILTRLR---LQKRGTGGVDTAAVGSVFDVSNADRLGSSEVEQVQ  351
            V L  L    + + ++  +R      RG  G  +    ++E VSN   LG +E E V+
Sbjct  169  VHLPGLVTTKRIKSVIEAIRSLGFVVRGIYGEGSMPASNIFQVSNQVTLGKTEAEIVE  226
```

Example: highly conserved gene families – creatine kinase

Sequences producing significant alignments:		Score (Bits)	E Value
qi|88958929|sp|Q724I1|Y243_LISMF	Hypothetical ATP:guanido phosph	69.3	2e-11
qi|25453326|sp|Q48759|Y231_LISMO	Hypothetical ATP:guanido phosph	68.9	2e-11
qi|25453355|sp|Q8XHP0|Y2442_CLOPE	Hypothetical ATP:guanido phosp	55.8	2e-07
qi|48428976|sp|P26460_1	[Segment 1 of 2] Creatine kinase B-t...	53.5	1e-06
qi|48428977|sp|P26460_2	[Segment 2 of 2] Creatine kinase B-t...	47.0	1e-04
qi|14916789|sp|Q9Z7K4|Y701_CHLPN	Hypothetical ATP:guanido pho...	36.6	0.14
qi|59797795|sp|Q72LR0|GIDA_LEPIC	tRNA uridine 5-carboxymethyl...	33.9	0.88

>qi|59797795|sp|Q72LR0|GIDA_LEPIC tRNA uridine 5-carboxymethylaminomethyl modification (Glucose-inhibited division protein A)
Length=635

Score = 33.9 bits (76), Expect = 0.88
Identities = 34/121 (28%), Positives = 54/121 (44%), Gaps = 30/121 (24%)

```
Query  29   HMAKVLTLELYKKLRDKETPSGFTVDDVIQTGVDNPGHPF------------------IM  70
            H  + LT  L+K+    E+  G  +DD++  GV++P   F                  +M
Sbjct  399  HSLRNLTPLLFKR---SESYIGVLIDDLVHKGVEDPYRMFTSRAEHRLLLRQDNADQRLM  455

Query  71   TVGCVAG--DEESYEVFKELFDPIISDRHGGY----KPTDKHKTDLNHE---NLKGGDDL  121
              G   G  D++SY+  KE ++ + S R   Y    KP+DK +  L+ +   N K G  L
Sbjct  456  KYGYDLGLVDQKSYDCMKEKYERVNSVREKIYQIPLKPSDKFQNLLDQKGITNYKFGMKL  515

Query  122  D  122
            D
Sbjct  516  D  516
```

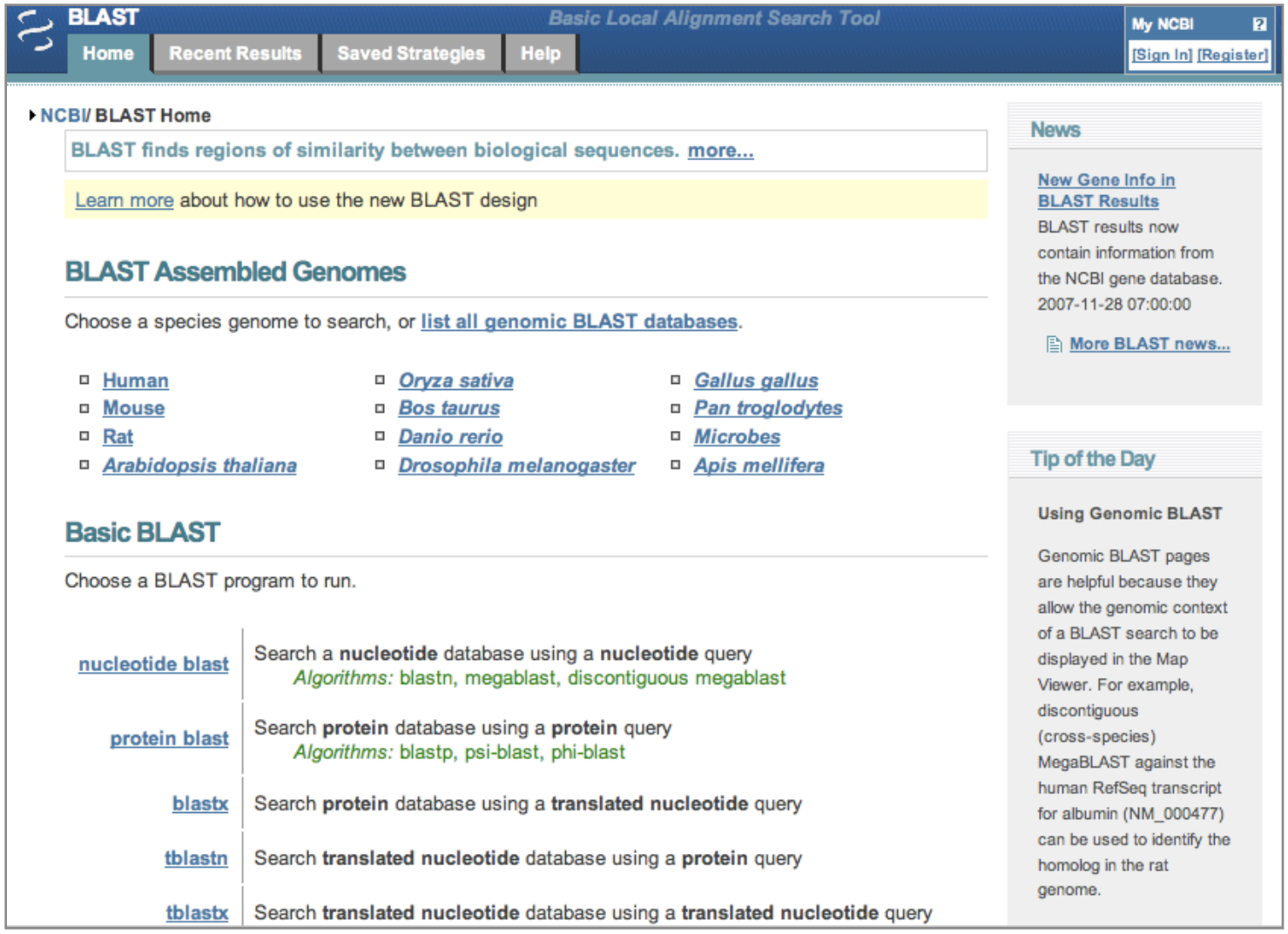
Example: poorly conserved homologs – haloacid dehalogenases

An important caveat - Statistical significance & biological significance are not necessarily the same thing!

```
Query= /phosphonatase/phosSt.gcg (255 letters) (10/20/99/pcb)
Database: /mol/seq/blast/db/swissprot 78,725 sequences; 28,368,147 total letters

                                                                    Score     E
Sequences producing significant alignments:                         (bits)  Value

sp|O06995|PGMB_BACSU Begin: 93 End: 204 PUTATIVE BETA-PHOSPHOGLUCOMUTASE (BETA-PGM)     38     0.020
sp|P31467|YIEH_ECOLI Begin: 1 End: 180 HYPOTHETICAL 24.7 KD PROTEIN IN TNAB-BGLB I...   36     0.10
sp|O14165|YDX1_SCHPO Begin: 34 End: 201 HYPOTHETICAL 27.1 KD PROTEIN C4C5.01 IN CHR...  31     2.6
sp|P41277|GPP1_YEAST Begin: 133 End: 200 (DL)-GLYCEROL-3-PHOSPHATASE 1                  30     4.4
sp|Q39565|DYHB_CHLRE Begin: 3911 End: 4032 DYNEIN BETA CHAIN, FLAGELLAR OUTER ARM       29     7.6
sp|P77625|YFBT_ECOLI Begin: 143 End: 187 HYPOTHETICAL 23.7 KD PROTEIN IN LRHA-ACKA I... 29     10.0
sp|Q40297|FCPA_MACPY Begin: 146 End: 176 FUCOXANTHIN-CHLOROPHYLL A-C BINDING PROTEIN... 29     13
sp|P40853|GPHP_ALCEU Begin: 94 End: 188 PHOSPHOGLYCOLATE PHOSPHATASE, PLASMID (PGP)     29     13
sp|Q40296|FCPB_MACPY Begin: 146 End: 176 FUCOXANTHIN-CHLOROPHYLL A-C BINDING PROTEIN... 29     13
sp|P52183|ANNU_SCHAM Begin: 119 End: 168 ANNULIN (PROTEIN-GLUTAMINE GAMMA-GLUTAMYLTR... 29     13
sp|P40106|GPP2_YEAST Begin: 133 End: 200 (DL)-GLYCEROL-3-PHOSPHATASE 2                  28     17
sp|P37934|MAY3_SCHCO Begin: 435 End: 552 MATING-TYPE PROTEIN A-ALPHA Y3                 27     29
sp|O06219|MURE_MYCTU Begin: 255 End: 371 UDP-N-ACETYLMURAMOYLALANYL-D-GLUTAMATE--2,6... 27     29
sp|P08419|EL2_PIG Begin: 182 End: 245 ELASTASE 2 PRECURSO                              27     38
sp|Q11034|Y07S_MYCTU Begin: 163 End: 218 HYPOTHETICAL 69.5 KD PROTEIN CY02B10.28C      27     38
sp|P00577|RPOC_ECOLI Begin: 1290 End: 1401 DNA-DIRECTED RNA POLYMERASE BETA' CHAIN (T  27     38
sp|P32662|GPH_ECOLI Begin: 20 End: 49 PHOSPHOGLYCOLATE PHOSPHATASE (PGP)               27     38
sp|P32662|GPH_ECOLI Begin: 116 End: 224 PHOSPHOGLYCOLATE PHOSPHATASE (PGP)             27     28
sp|P32282|RIR1_BPT4 Begin: 239 End: 266 RIBONUCLEOSIDE-DIPHOSPHATE REDUCTASE ALPHA C... 27     50
sp|P17346|LEC2_MEGRO Begin: 36 End: 121 LECTIN BRA-2                                   27     50
sp|P54947|YXEH_BACSU Begin: 24 End: 51 HYPOTHETICAL 30.2 KD PROTEIN IN IDH-DEOR IN...  27     50
sp|P77366|PGMB_ECOLI Begin: 95 End: 190 PUTATIVE BETA-PHOSPHOGLUCOMUTASE (BETA-PGM)    27     50
sp|P30139|THIG_ECOLI Begin: 43 End: 79 THIG PROTEIN                                    27     50
sp|P95649|CBBY_RHOSH Begin: 96 End: 189 CBBY PROTEIN                                   27     50
sp|Q43154|GSHC_SPIOL Begin: 228 End: 327 GLUTATHIONE REDUCTASE, CHLOROPLAST PRECURSO... 26     66
sp|P34132|NT6A_HUMAN Begin: 191 End: 215 NEUROTROPHIN-6 ALPHA (NT-6 ALPHA)             26     66
sp|P34134|NT6G_HUMAN Begin: 115 End: 144 NEUROTROPHIN-6 GAMMA (NT-6 GAMMA)             26     66
sp|P95650|GPH_RHOSH Begin: 48 End: 114 PHOSPHOGLYCOLATE PHOSPHATASE (PGP)              26     66
```

BLAST

(see http://www.ncbi.nlm.nih.gov/BLAST/ for excellent documentation)



Extending our reach: Psi-Blast, etc.

(Altschul et al., "Gapped Blast and PSI-Blast: A new generation of protein database search programs," NAR 25:3389-3402, 1997)

- Generalizes BLAST algorithm to use a position-specific score matrix in place of a query sequence & associated substitution matrix for searching the databases
- Position-specific score matrix generated from the output of a gapped Blast search, *i.e.,* uses a profile or motif defined in the initial Blast search in place of a single query sequence and matrix for subsequent searches of the database
- Results in a database search "tuned" to the specific sequence characteristics of interest
- Good at finding remote homologs but contains no clustering information

Evaluation of sequence relationships using multiple alignments

Multiple alignments provide more information than pairwise alignments

- Identification of conserved elements important to function
- Determination of the level and sites of variability across the members of subgroups/families/ superfamilies
- Many tools & compilations of pre-computed alignments available

Example: poorly conserved homologs – haloacid dehalogenases

```
BLASTP 2.0a19MP-WashU [05-Feb-1998] [Build decunix3.2 01:53:21 05-Feb-1998]

Query= /phosphonatase/phosBc.gcg        (302 letters)

Database:  swissprot
           77,273 sequences; 27,815,109 total letters.

                                                                  Smalle
                                                                    Sum
                                                          High   Probabi
Sequences producing High-scoring Segment Pairs:           Score   P(N)

sp|P77247|YNIC_ECOLI HYPOTHETICAL 24.3 KD PROTEIN IN PFKB...   116  2.2e-05
sp|O67359|GPH_AQUAE  PHOSPHOGLYCOLATE PHOSPHATASE (PGP)        106  0.00030
sp|O06995|PGMB_BACSU PUTATIVE BETA-PHOSPHOGLUCOMUTASE (BE...    97  0.0039
sp|P31467|YIEH_ECOLI HYPOTHETICAL 24.7 KD PROTEIN IN TNAB...    94  0.0082
sp|P44755|GPH_HAEIN  PHOSPHOGLYCOLATE PHOSPHATASE (PGP)         93  0.011
sp|P54607|YHCW_BACSU HYPOTHETICAL 24.7 KD PROTEIN IN CSPB...    89  0.030
sp|P32662|GPH_ECOLI  PHOSPHOGLYCOLATE PHOSPHATASE (PGP)         87  0.067
```

```
PGPhos     -----MPGVVFDLDGTLVHSAPDIHAAVNK
Phosphon   MDRMKIEAVIFDWAGTTVDYGCFAPLEVFM

PGPhos     ALAEEGGAPFTLAEITGFIG--NGVPVLIQ
Phosphon   EIFHKRGVAITAEEARKPMGLLKIDHVRVT

PGPhos     RVLAARGEAPDAHRQAELQGRFMAHYEADP
Phosphon   EMPRIASEWNRVFRQLPTEADIQEMYEEFE

PGPhos     ATLTSVYP-------GAEAAIRHLRAEGWR
Phosphon   EILFAILPRYASPINGVKEVIASLRERGIK

PGPhos     IGLCTNKPVGASRQILSLF---GLLELF--
Phosphon   IGSTT----GYTMEMMDIVAKEAALQGYKP

PGPhos     DAIIGGESLPQRKPDPAPLRATAAALN---
Phosphon   DFLVTPDDVPAGRPYPWMSYKNAMELGVYP

PGPhos     EEVVLYVGDSEVDAATAEAAGLRFALFTEG
Phosphon   MNHMIKVGDTVSDMKEGRNAGMWTVGVILG

PGPhos     YRHAPV--HELPHHGLFSHHDELQDLLRRL
Phosphon   SSELGLTEEEVENMDSVELREKIEVVRNRF
```

~ 21% identical

```
                10          151         176
                *           *           *
Cu++ATPase.Ec   LDTVVFDKTGTLTEG VIAGVLPDGKAEAIKHL AMVGDGINDAPAL
Cu++ATPase.Hs   VKVVVFDKTGTITHG VFAEVLPSHKVAKVKQL AMVGDGINDSPAL
Ca++ATPase.At   ATTICSDKTGTLTTN VMARSSPMDKHTLVRLL AVTGDGTNDAPAL
Urf.Mj          KVAIVFDSAGTLVKI E---AHQELKRDLIRNL IMVGDGANDVPAM
PhosSerPhos.Hs  ADAVCFDVDSTVIRE TAE-SGGKGKVIKLLKE IMIGDGATDMEAC
2-DO-6-PPhos.Sc VDLCLFDLDGTIVST ITGFDVKNGKPDPEGYS VVFEDAPVGIKAG
DL-Gly-3-Phos.Sc INAALFDVDGTIIIS ITANDVKQGKPHPEPYL VVFEDAPAGIAAG
Phosphon.Pa     LQAAILDWAGTVVDF ATDEV-PNGRPWPAQAL VKVDDTWPGILEG
Phosphon.St     IHAVILDWAGTTVDF ATDDLAAGGRPGPWMAL VKVDDAAPGISEG
Phosphon.Bc     IEAVIFDWAGTTVDY TPDDV-PAGRPYPWMSY IKVGDTVSDMKEG
PhosGlycolPhos.Rs MPGVVFDLDGTLVHS IGGESLPQRKPDPAPLA LYVGDSEVDAATA
NtermDom.IGPD.Pp VQALLLDMDGVMAEV LEDCPP---KPSPEPIL AMVGDTVDDIIAG
B-PhosGlucoMut.Ll FKAVLFDLDGVITDT AEVAAS---KPAPDIFI IGLEDSQAGIQAI
HaloAcidDehal.PspYL IKGIAFDLYGTLFDV LSVDPVQVYKPDNRVYE LFVSSNAWDATGA
NtermDomEpoxHyd.Hs LRAAVFDLDGVLALP IESCQVGMVKPEPQIYK VFLDDIGANLKPA
EnolasePhos.Ko  IRAIVTDIEGTTSDI FD--TLVGAKREAQSYR LFLSDIHQELDAA
```

Example: highly conserved gene families – creatine kinase

What is the range of divergence among the sequences you plan to align?

Sequences producing significant alignments:		Score (Bits)	E Value
qi|125305|sp|P06732|KCRM_HUMAN	Creatine kinase M-type (Creatine	757	0.0
qi|125303|sp|P05123|KCRM_CANFA	Creatine kinase M-type (Creatine	740	0.0
qi|62286641|sp|Q5XLD3|KCRM_PIG	Creatine kinase M-type (Creatine	739	0.0
⋮			
qi|125304|sp|P00565|KCRM_CHICK	Creatine kinase M-type (Creatine	695	0.0
qi|125310|sp|P00566|KCRM_TORMA	Creatine kinase M-type (Creati...	665	0.0
qi|125309|sp|P04414|KCRM_TORCA	Creatine kinase M-type (Creatine	664	0.0

pairwise % identity:

Human CKM vs Torpedo california CKM = 84%

Example: highly conserved gene families – creatine kinase

Alignment of highly similar proteins

Example: highly conserved gene families – creatine kinase

What is the range of divergence among the sequences you plan to align?

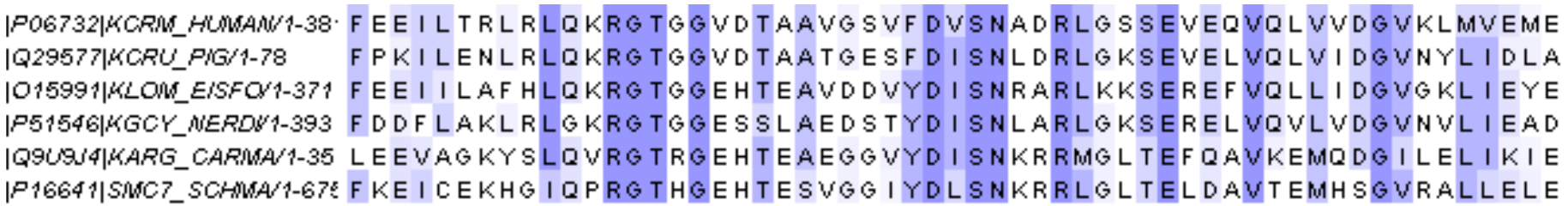
Sequences producing significant alignments:		Score (Bits)	E Value
gi|125305|sp|P06732|KCRM_HUMAN	Creatine kinase M-type (Creatine	748	0.0
⋮			
gi|1730042|sp|P51546|KGCY_NERDI	Glycocyamine kinase (GK) (Guanid	421	3e-117
⋮			
gi|3183058|sp|O15991|KLOM_EISFO	Lombricine kinase (LK)	370	4e-102
⋮			
gi|25453075|sp|Q9U9J4|KARG_CARMA	Arginine kinase (AK)	231	4e-60
⋮			
gi|134552|sp|P16641|SMC7_SCHMA	ATP:guanidino kinase SMC74 (ATP:g	191	4e-48
⋮			
gi|3183053|sp|Q29577|KCRU_PIG	Creatine kinase, ubiquitous mit...	114	7e-25

pairwise % identity:

Human CKM vs Schistosoma guanidino kinase= 36%

Example: highly conserved gene families – creatine kinase

Alignment of highly similar proteins



…and less similar proteins that are still homologous

Example: poorly conserved homologs – haloacid dehalogenases

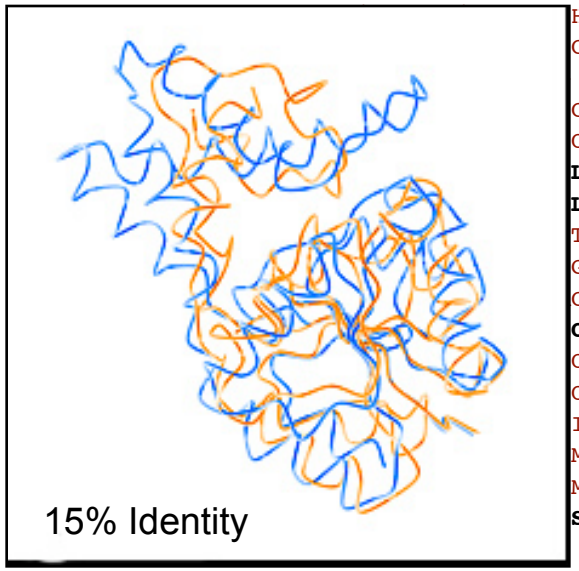
Adding structural information

```
Query= /phosphonatase/phosSt.gcg (255 letters) (10/20/99/pcb)
Database: /mol/seq/blast/db/swissprot 78,725 sequences; 28,368,147 total letters

                                                                    Score     E
Sequences producing significant alignments:                         (bits)  Value

sp|O06995|PGMB_BACSU Begin: 93 End: 204 PUTATIVE BETA-PHOSPHOGLUCOMUTASE (BETA-PGM)    38     0.020
sp|P31467|YIEH_ECOLI Begin: 1 End: 180 HYPOTHETICAL 24.7 KD PROTEIN IN TNAB-BGLB I...  36     0.10
sp|O14165|YDX1_SCHPO Begin: 34 End: 201 HYPOTHETICAL 27.1 KD PROTEIN C4C5.01 IN CHR...  31     2.6
sp|P41277|GPP1_YEAST Begin: 133 End: 200 (DL)-GLYCEROL-3-PHOSPHATASE 1                  30     4.4
sp|Q39565|DYHB_CHLRE Begin: 3911 End: 4032 DYNEIN BETA CHAIN, FLAGELLAR OUTER ARM      29     7.6
sp|P77625|YFBT_ECOLI Begin: 143 End: 187 HYPOTHETICAL 23.7 KD PROTEIN IN LRHA-ACKA I...  29     10.0
sp|Q40297|FCPA_MACPY Begin: 146 End: 176 FUCOXANTHIN-CHLOROPHYLL A-C BINDING PROTEIN...  29     13
sp|P40853|GPHP_ALCEU Begin: 94 End: 188 PHOSPHOGLYCOLATE PHOSPHATASE, PLASMID (PGP)    29     13
sp|Q40296|FCPB_MACPY Begin: 146 End: 176 FUCOXANTHIN-CHLOROPHYLL A-C BINDING PROTEIN...  29     13
sp|P52183|ANNU_SCHAM Begin: 119 End: 168 ANNULIN (PROTEIN-GLUTAMINE GAMMA-GLUTAMYLTR...  29     13
sp|P40106|GPP2_YEAST Begin: 133 End: 200 (DL)-GLYCEROL-3-PHOSPHATASE 2                  28     17
sp|P37934|MAY3_SCHCO Begin: 435 End: 552 MATING-TYPE PROTEIN A-ALPHA Y3                 27     29
sp|O06219|MURE_MYCTU Begin: 255 End: 371 UDP-N-ACETYLMURAMOYLALANYL-D-GLUTAMATE--2,6...  27     29
sp|P08419|EL2_PIG Begin: 182 End: 245 ELASTASE 2 PRECURSO                           27     38
sp|Q11034|Y07S_MYCTU Begin: 163 End: 218 HYPOTHETICAL 69.5 KD PROTEIN CY02B10.28C     27     38
sp|P00577|RPOC_ECOLI Begin: 1290 End: 1401 DNA-DIRECTED RNA POLYMERASE BETA' CHAIN (T  27     38
sp|P32662|GPH_ECOLI Begin: 20 End: 49 PHOSPHOGLYCOLATE PHOSPHATASE (PGP)            27     38
sp|P32662|GPH_ECOLI Begin: 116 End: 224 PHOSPHOGLYCOLATE PHOSPHATASE (PGP)          27     28
sp|P32282|RIR1_BPT4 Begin: 239 End: 266 RIBONUCLEOSIDE-DIPHOSPHATE REDUCTASE ALPHA C...  27     50
sp|P17346|LEC2_MEGRO Begin: 36 End: 121 LECTIN BRA-2                                27     50
sp|P54947|YXEH_BACSU Begin: 24 End: 51 HYPOTHETICAL 30.2 KD PROTEIN IN IDH-DEOR IN...  27     50
sp|P77366|PGMB_ECOLI Begin: 95 End: 190 PUTATIVE BETA-PHOSPHOGLUCOMUTASE (BETA-PGM)    27     50
sp|P30139|THIG_ECOLI Begin: 43 End: 79 THIG PROTEIN                                 27     50
sp|P95649|CBBY_RHOSH Begin: 96 End: 189 CBBY PROTEIN                                27     50
sp|Q43154|GSHC_SPIOL Begin: 228 End: 327 GLUTATHIONE REDUCTASE, CHLOROPLAST PRECURSO...  26     66
sp|P34132|NT6A_HUMAN Begin: 191 End: 215 NEUROTROPHIN-6 ALPHA (NT-6 ALPHA)          26     66
sp|P34134|NT6G_HUMAN Begin: 115 End: 144 NEUROTROPHIN-6 GAMMA (NT-6 GAMMA)          26     66
sp|P95650|GPH_RHOSH Begin: 48 End: 114 PHOSPHOGLYCOLATE PHOSPHATASE (PGP)           26     66
```

Example: poorly conserved homologs – haloacid dehalogenases

Adding structural information

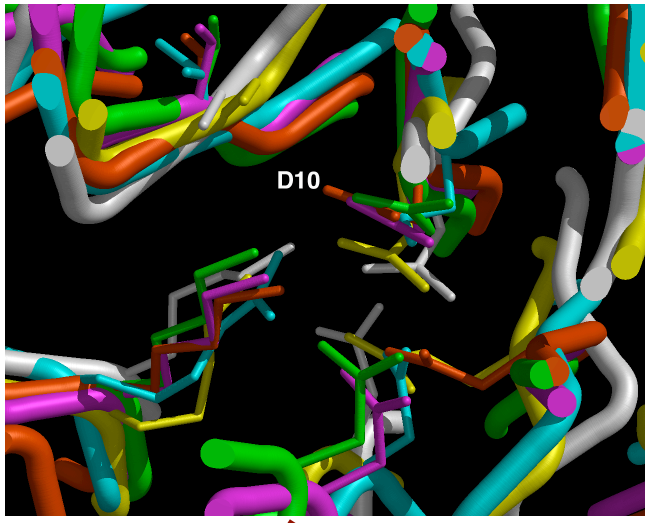
```
Query= /phosphonatase/phosSt.gcg (255 letters) (10/20/99/pcb)
Database: /mol/seq/blast/db/swissprot 78,725 sequences; 28,368,147 total letters

                                                                    Score     E
Sequences producing significant alignments:                         (bits)  Value

sp|O06995|PGMB_BACSU Begin: 93 End: 204 PUTATIVE BETA-PHOSPHOGLUCOMUTASE (BETA-PGM)    38     0.020
sp|P31467|YIEH_ECOLI Begin: 1 End: 180 HYPOTHETICAL 24.7 KD PROTEIN IN TNAB-BGLB I...  36     0.10
sp|O14165|YDX1_SCHPO Begin: 34 End: 201 HYPOTHETICAL 27.1 KD PROTEIN C4C5.01 IN CHR...  31     2.6
sp|P41277|GPP1_YEAST Begin: 133 End: 200 (DL)-GLYCEROL-3-PHOSPHATASE 1              30     4.4
sp|Q39565|DYHB_CHLRE Begin: 3911 End: 4032 DYNEIN BETA CHAIN, FLAGELLAR OUTER ARM   29     7.6
sp|P77625|YFBT_ECOLI Begin: 143 End: 187 HYPOTHETICAL 23.7 KD PROTEIN IN LRHA-ACKA I...  29     10.0
sp|Q40297|FCPA_MACPY Begin: 146 End: 176 FUCOXANTHIN-CHLOROPHYLL A-C BINDING PROTEIN...  29     13
sp|P40853|GPHP_ALCEU Begin: 94 End: 188 PHOSPHOGLYCOLATE PHOSPHATASE, PLASMID (PGP)  29     13
sp|Q40296|FCPB_MACPY Begin: 146 End: 176 FUCOXANTHIN-CHLOROPHYLL A-C BINDING PROTEIN...  29     13
sp|P52183|ANNU_SCHAM Begin: 119 End: 168 ANNULIN (PROTEIN-GLUTAMINE GAMMA-GLUTAMYLTR...  29     13
sp|P40106|GPP2_YEAST Begin: 133 End: 200 (DL)-GLYCEROL-3-PHOSPHATASE 2              28     17
HCO Begin: 435 End: 552 MATING-TYPE PROTEIN A-ALPHA Y3                              27     29
CTU Begin: 255 End: 371 UDP-N-ACETYLMURAMOYLALANYL-D-GLUTAMATE--2,6...              27     29
 Begin: 182 End: 245 ELASTASE 2 PRECURSO                                            27     38
CTU Begin: 163 End: 218 HYPOTHETICAL 69.5 KD PROTEIN CY02B10.28C                    27     38
OLI Begin: 1290 End: 1401 DNA-DIRECTED RNA POLYMERASE BETA' CHAIN (T                27     38
LI Begin: 20 End: 49 PHOSPHOGLYCOLATE PHOSPHATASE (PGP)                             27     38
LI Begin: 116 End: 224 PHOSPHOGLYCOLATE PHOSPHATASE (PGP)                           27     28
T4 Begin: 239 End: 266 RIBONUCLEOSIDE-DIPHOSPHATE REDUCTASE ALPHA C...              27     50
GRO Begin: 36 End: 121 LECTIN BRA-2                                                 27     50
CSU Begin: 24 End: 51 HYPOTHETICAL 30.2 KD PROTEIN IN IDH-DEOR IN...                27     50
OLI Begin: 95 End: 190 PUTATIVE BETA-PHOSPHOGLUCOMUTASE (BETA-PGM)                  27     50
OLI Begin: 43 End: 79 THIG PROTEIN                                                  27     50
OSH Begin: 96 End: 189 CBBY PROTEIN                                                 27     50
IOL Begin: 228 End: 327 GLUTATHIONE REDUCTASE, CHLOROPLAST PRECURSO...              26     66
MAN Begin: 191 End: 215 NEUROTROPHIN-6 ALPHA (NT-6 ALPHA)                           26     66
MAN Begin: 115 End: 144 NEUROTROPHIN-6 GAMMA (NT-6 GAMMA)                           26     66
SH Begin: 48 End: 114 PHOSPHOGLYCOLATE PHOSPHATASE (PGP)                            26     66
```

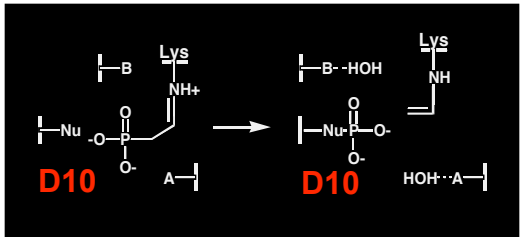


Confirmation of relationships for biological insight

› map conserved elements of structure to conserved elements of function

10	151	180
F D K T G T	K A	G D G I N D
F D K T G T	K V	G D G I N D
S D K T G T	K H	G D G T N D
F D S A G T	K R	G D G A N D
F D V D S T	K V	G D G A T D
F D L D G T	K P	E D A P V G
F D V D G T	K P	E D A P A G
L D W A G T	R P	D D T W P G
L D W A G T	R P	D D A A P G
F D W A G T	R P	G D T V S D
F D L D G T	K P	G D S E V D
L D M D G V	K P	G D T V D D
F D L D G V	K P	E D S Q A G
F D L Y G T	K P	S S N A W D
F D L D G V	K P	D D I G A N
T D I E G T	K R	S D I H Q E





› experiment!

www.rcsb.org/pdb

RCSB PDB PROTEIN DATA BANK

A MEMBER OF THE PDB

An Information Portal to Biological Macromolecular Structures

As of Tuesday Mar 25, 2008 there are 49760 Structures | PDB Statistics

CONTACT US | HELP | PRINT PAGE PDB ID or keyword Author Site Search | Advanced Search

Home | Search | Structure | Results | Queries

- 1I0E
- Download Files
- FASTA Sequence
- Download Original Files
- Display Files
- Display Molecule
- Structural Reports
- External Links
- Structure Analysis
- Help

Quick Tips :

Click the PDB file icon above to view the PDB file.

Are you missing data updates? The PDB archive has moved to ftp://ftp.wwpdb.org. For more information click here.

Help | Structure Summary | Biology & Chemistry | Materials & Methods | Sequence Details | Geometry

1i0e

DOI 10.2210/pdb1i0e/pdb

Red - Derived Information

Title	CRYSTAL STRUCTURE OF CREATINE KINASE FROM HUMAN MUSCLE
Authors	Shen, Y.-Q., Tang, L., Zhou, H.-M., Lin, Z.-J.
Primary Citation	Shen, Y.Q., Tang, L., Zhou, H.M., Lin, Z.J. (2001) Structure of human muscle creatine kinase. *Acta Crystallogr.,Sect.D* ***57:*** *1196-1200* [Abstract]
History	Deposition 2001-01-29 Release 2003-04-01
Experimental Method	Type X-RAY DIFFRACTION Data [EDS]

Resolution[Å] R-Value R-Free Space Group

Images and Visualization

<< Biological Molecule 1 >>

Display Options

KiNG

Jmol

WebMol

Resources

Acknowledgements

Patricia C. Babbitt

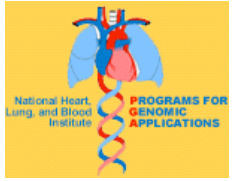
Courtney Harper

Ranyee Chiang

Susan Johns

NATIONAL HEART, LUNG, AND BLOOD INSTITUTE

Programs for Genomic Applications



Acknowledgements

Patricia C. Babbitt

Courtney Harper

Ranyee Chiang

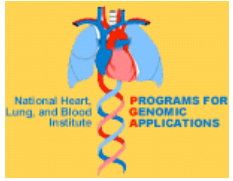
Susan Johns

Lecture & Workshop available online

(http://www.rbvi.ucsf.edu/Outreach/Workshops/BayGenomics_traveling_tutorial/seqanalysis.html)

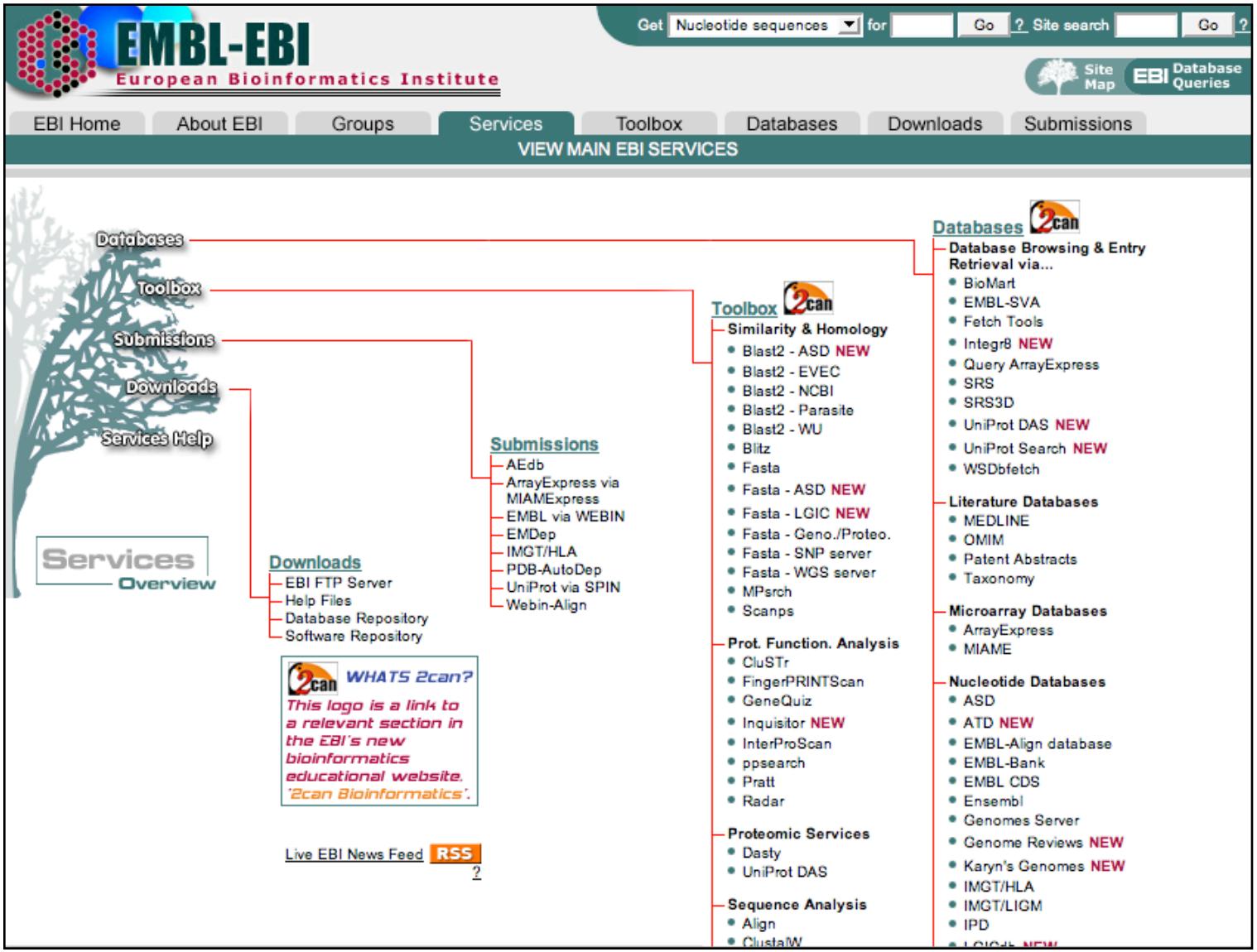
NATIONAL HEART, LUNG, AND BLOOD INSTITUTE

Programs for Genomic Applications

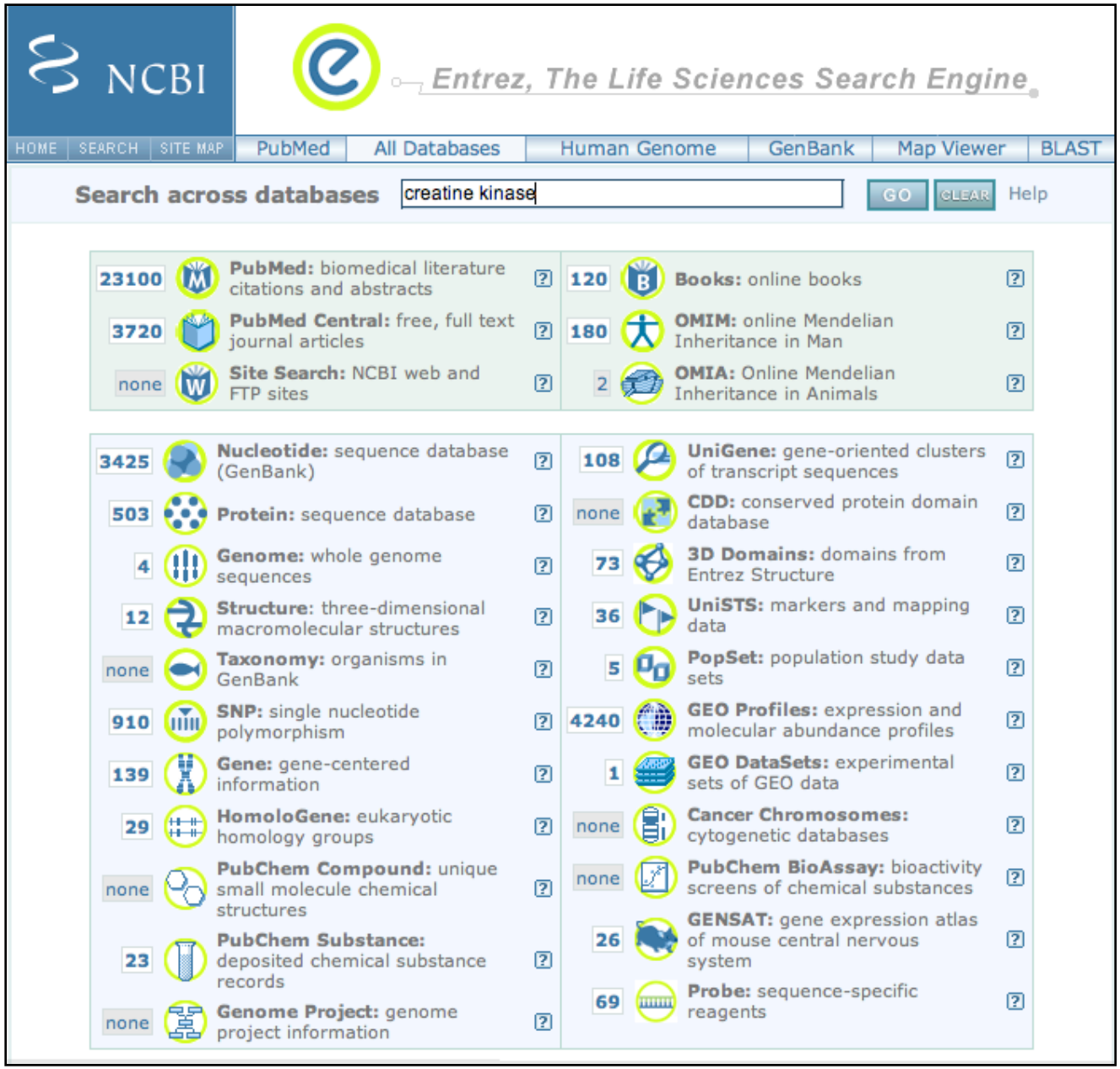


One-stop shopping

http://193.62.197.151/services/services_tree.html



see http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi



Primary sequence/structure resources

Entrez: Access site for multiple types of databases at NIH NCBI, including GenBank, GenPept, Gene, Genomces, SNPs, etc.
http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi

SwissProt: A highly curated protein sequence database
http://www.expasy.org/sprot/

Pfam: Pre-computed library of multiple sequence alignments and HMM*s for many protein domains (7316 families as of 1/04) http://pfam.wustl.edu/

Superfamily: Pre-computed multiple sequence alignments & HMMs based on 1539 structural superfamilies associated with the SCOP hierarchy (see below), http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/

Protein Data Bank: The repository of all publicly available 3D structures
http://www.rcsb.org/pdb/Welcome.do

SCOP (Structural Classification of Proteins): Compilation of publicly available 3D structures organized hierarchically by fold, superfamily, family http://scop.mrc-lmb.cam.ac.uk/scop/

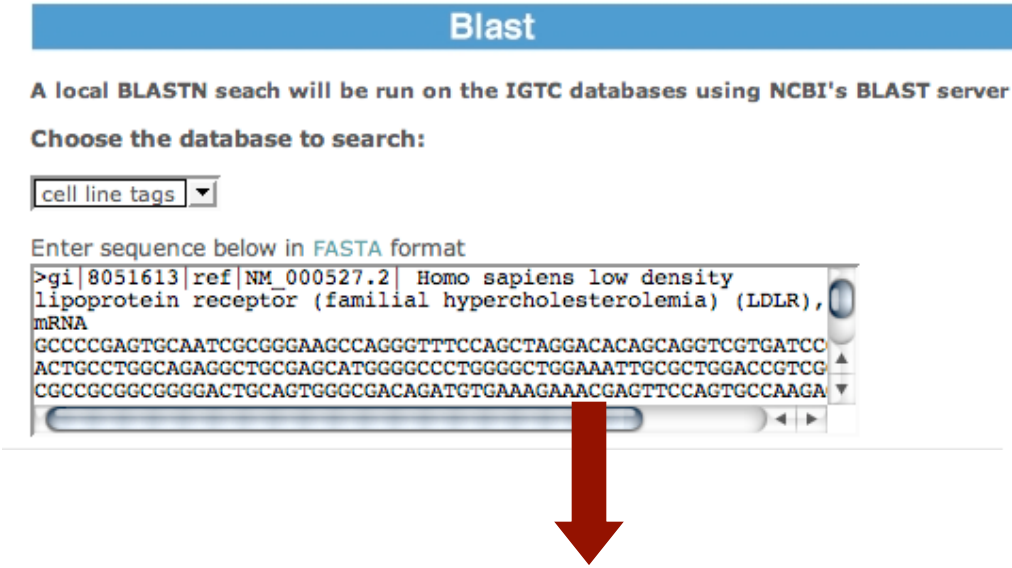
MODBASE: A database of structural models (3,094,524 reliable models or PSI-BLAST fold assignments for domains in 1,094,750 proteins)
http://modbase.compbio.ucsf.edu/modbase-cgi-new/search_form.cgi

*HMM = Hidden Markov model, as used for sequence analysis: A probabilistic model used to align and analyze sequence datasets by generalization from a sequence profile

Tools

Note: Searching any genomics database with either a nucleic acid or protein sequence requires specific formats that the resource software can recognize

- Most common: Fasta format (for more information, see http://www.ncbi.nlm.nih.gov/BLAST/fasta.shtml)
- Many others, including plain text



>gi|8051613|ref|NM_000527.2| Homo sapiens low density lipoprotein receptor (familial hypercholesterolemia) (LDLR), mRNA

GCCCCGAGTGCAATCGCGGGAAGCCAGGGTTTCCAGCTAGGACACAGCAGGTCGTGATCCGGGTCGGGAC
ACTGCCTGGCAGAGGCTGCGAGCATGGGGCCCTGGGGCTGGAAATTGCGCTGGACCGTCGCCTTGCTCCT
CGCCGCGGCGGGGACTGCAGTGGGCGACAGATGTGAAAGAAACGAGTTCCAGTGCCAAGACGGGAAATGC
ATCTCCTACAAGTGGGTCTGCGATGGCAGCGCTGAGTGCCAGGATGGCTCTGATGAGTCCCAGGAGACGT
GCTTGTCTGTCACCTGCAAATCCGGGGACTTCAGCTGTGGGGGCCGTGTCAACCGCTGCATTCCTCAGTT

Multiple alignment tools

- Local alignment methods identify ordered series of motifs, then aligns the intervening regions
  - Examples: MACAW, PIMA
- Global alignment methods construct an alignment throughout the length of the entire sequence
  - Examples: Pileup, Clustal family, MSA
- HMMs & profile methods generate a profile or position-specific scoring system based on the liklihood of seeing a particular residue at a given position
- T-Coffee attempts to combine the best of both local and global alignments
- MUSCLE combines tree-building with progressive global alignments & profile methods - fast becoming the method of choice in the field

Tools for working with 3D structures

Structure comparison

- Fold-based
  - Combinatorial Extension algorithm (CE); server available at http://cl.sdsc.edu/
  - Dali algorithm (pre-computed comparisons in FSSP DB); server at http://www.ebi.ac.uk/dali/
- Active-site templates
  - Fuzzy Functional forms, SPASM, JESS

Structure visualization/mapping to sequence info

- Evolutionary Trace (no really good automated servers yet available)
- Chimera - structural visualization software developed at UCSF, free download at http://www.cgl.ucsf.edu/chimera/

Multiple alignment program: ClustalW*

- Most commonly used implementation in a family of programs using profile-based progressive alignment
- Access: http://www2.ebi.ac.uk/clustalw/
- Permits user adjustment of many parameters for both the pairwise and multiple alignment stages
- Computes position-specific gap opening and extension penalties as the alignment proceeds, *e.g.,* varies parameters at different positions
- Uses different weight matrices for different alignments
- Allows addition of new sequences to an existing alignment without recomputing the entire alignment (recomputes weights)
- Alternate version, ClustalX, provides a windows interface for ClustalW

*"W" stands for "weighting" the sequences to correct for unequal sampling of sequences from different evolutionary distances

Evolutionary Trace

Takes advantage of the larger context provided by a family-based view of proteins to improve the accuracy of binding site determination

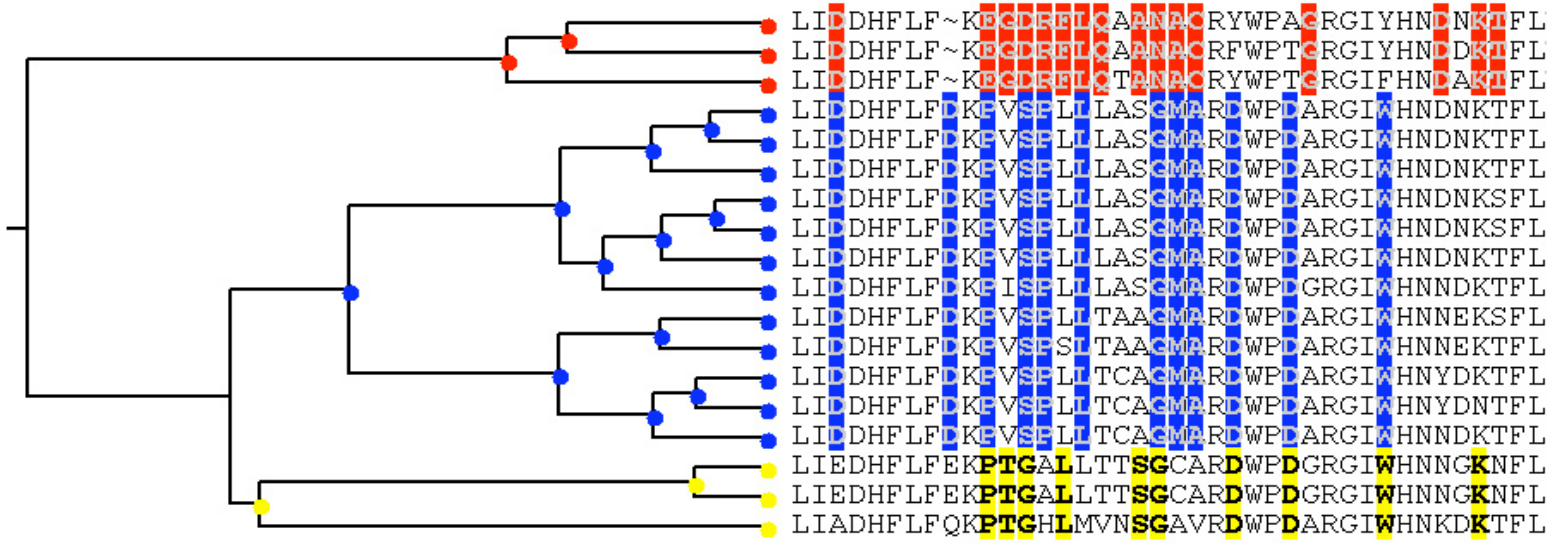
Uses sequence-based clustering of related proteins to distinguish class-specific differences in ligand binding determinants across a particular family or superfamily of proteins.

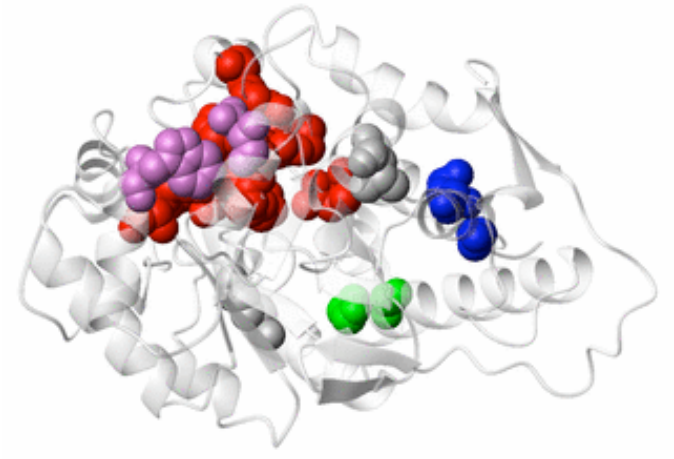
The sequence information derived from multiple alignments and phylogenetic clustering is leverage by mapping class specific patterns likely to be functionally important onto three dimensional structures and models

see papers by Lichtarge, O and Joachimiak MP & Cohen FE, “JEvTrace: refinement and variations of the evolutionary trace in JAVA,” Genome Biol. 3:(12), Epub 2002 Nov. 26

1. Select nodes

2. Select conserved or class conserved regions



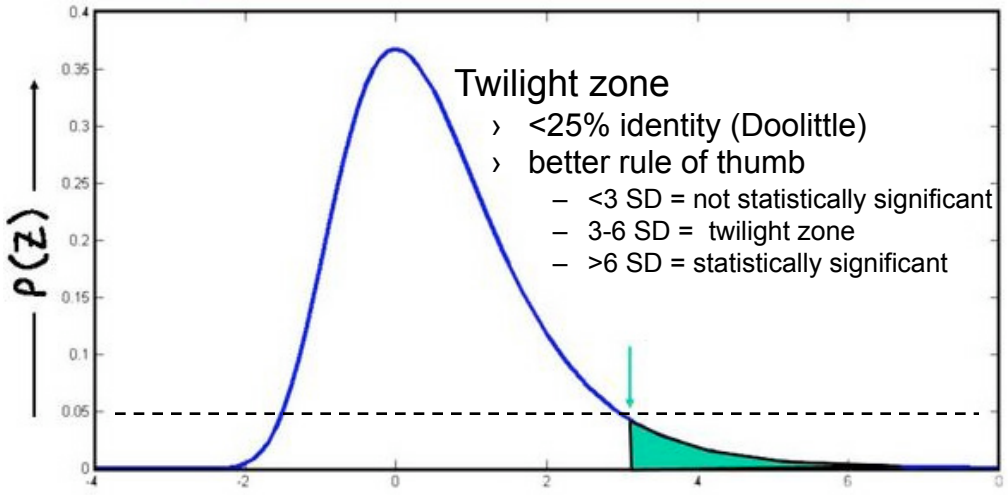


3. Map selections to structure

Statistical Significance

For searching large DBs, the ONLY criteria available to judge the likelihood of an evolutionary relationship between 2 sequences is an estimate of statistical significance

- to determine if the alignment score has statistical meaning, compare it with the score generated from the alignment of random sequences
- a model of random sequences is needed
  - simplest: choose the amino acid residues in a sequence independently, with background probabilities



References

Molecular Evolution (courtesy Margaret Glasner, PhD)

Web site

Workshop on Molecular Evolution (Marine Biological Laboratory, Woods Hole, MA)
http://workshop.molecularevolution.org/

Books

Li, Wen-Hsiung (1997) *Molecular Evolution*. Sinauer Associates.

Nei, Masatoshi and Kumar, Sudhir (2000) *Molecular Evolution and Phylogenetics*. Oxford University Press.

Graur, Dan and Li, Wen-Hsiung (2000) *Fundamentals of Molecular Evolution*, 2nd Edition. Sinauer Associates.

Hall, Barry (2004) *Phylogenetic Trees Made Easy: A How-To Manual*, 2nd Edition. Sinauer Associates. (This is a great book that introduces several phylogenetic software packages.)

Swofford, D. L., G. J. Olsen, P. J. Waddell and D. M. Hillis. 1996. Phylogenetic inference. Pp. 407-543 in D. M. Hillis, C. Moritz and B. K. Mable, eds., Molecular Systematics, second edition. Sinauer Associates, Sunderland, Massachusetts. (This is the only useful chapter in this book.)

Papers

Ronquist and Huelsenbeck (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics 19:1572-1574.*

Bielawski, J.P., and Yang, Z. (2003) Maximum likelihood methods for detecting adaptive evolution after gene duplication. *Journal of Structural and Functional Genomics* 3: 201–212.

Jermann, T. M., Opitz, J. G., Stackhouse, J., and Benner, S. A. (1995) Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature* 374:57-59.

Bielawski JP, Dunn KA, Sabehi G, Beja O. (2004) Darwinian adaptation of proteorhodopsin to different light intensities in the marine environment. *Proc Natl Acad Sci U S A* 101:14824-9.

PAML (Phylogenetic Analysis Using Maximum Likelihood) **Software for studying adaptive evolution**

http://abacus.gene.ucl.ac.uk/software/paml.html

Sequence Analysis - Seminal papers

Needleman & Wunsch, “A general method applicable to the search for similarities in the amino acid sequence of two proteins,” JMB 48: 443-453, 1970

Smith & Waterman, “Identification of Common Molecular Subsequences,” JMB 147:195-197, 1981

Pearson & Lipman, (the original FASTA paper) “Improved tools for biological sequence comparison,” PNAS USA 80:1382-1386, 1983

Gribskov et al., “Profile analysis: Detection of distantly related proteins,” PNAS USA 84: 4355-4358, 1987

Altschul et al., (the original Blast paper) "Basic local alignment search tool,” JMB 215;403-410, 1990

Pascarella & Argos, “Analysis of Insertions/Deletions in Protein Structures,” JMB. 224, 461-471, 1992

Henikoff & Henikoff, “Amino acid substitution matrices from protein blocks,” PNAS USA 89:10915-10919, 1992

Altschul et al., “Gapped Blast and PSI-Blast: A new generation of protein database search programs,” NAR 25:3389-3402, 1997

Yona G, Levitt M “Within the twilight zone: a sensitive profile-profile comparison tool based on information theory,” JMB 315: 1257-1275, 2002

Sadreyev R, Grishin N “COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance,” JMB 326: 317-336, 2003

John B, Sali A “Detection of homologous proteins by an intermediate sequence search,” Prot. Sci. 13: 54-62, 2004

Multiple Alignments

Feng, DF & Doolittle, RF, “Progressive sequence alignment as a prerequisite to correct phylogenetic trees,” J Mol Evol 25:351-360, 1987 (PILEUP)

Smith, RF & Smith, TF “Automatic generation of primary sequence patterns from sets of related protein sequences,” PNAS USA 87:118-122, 1990 (PIMA)

Thompson, JD et al., “ClustalW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalities and weight matrix choice,” NAR 22:4673-4680, 1994

Eddy, S “Multiple alignment using hidden Markov models,”Proc Int Conf Intell Syst Mol Biol. 3:114-20, 1995

Notredame, C & Higgins, DG, “SAGA: Sequence alignment by genetic algorithm,” NAR 24:1515-1524, 1996

Eddy, S “Profile hidden Markov models,” Bioinf. 14:755-763, 1998

Notredame et al, “T-Coffee: A novel method for fast and accurate multiple sequence alignment,” JMB 302: 205-217, 2000

Tang C et al, “On the role of structural information in remote homology detection and sequence alignment: New methods using hybrid sequence profiles,” JMB 334: 1043-1062, 2003

Edgar, RC “MUSCLE: a multiple sequence alignment method with reduced time and space complexity,” BMC Bioinformatics 19:113, 2004

A few more seminal papers in computational biology

Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. EMBO J 5, 823-826 (1986)

Ponder JW, Richards FM. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes.J Mol Biol. 193, 775-91 (1987)

Sippl, M. J. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 213, 859-83 (1990)

S. Pascarella and P. Argos, "Analysis of Insertions/Deletions in Protein Structures", *J. Mol. Biol*. 224, 461-471 (1992)

C.A. Orengo, D.T. Jones and J.M Thornton, "Protein Superfamilies and Domain Superfolds", *Nature* 372**,** 631-634 (1994)

O. Lichtarge, H.R. Bourne and F.E. Cohen, "An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families", *J. Mol. Biol*. 257, 342-358 (1996)

A.J. Enright, I. Iliopoulos, N.C. Kyrpides, C.A. Ouzounis, "Protein interaction maps for complete genomes based on gene fusion events," *Nature* 402, 86-90 (1999)

Baker, D. A surprising simplicity to protein folding. Nature 405, 39-42 (2000)

Victor Neduva, Rune Linding, Isabelle Su-Angrand, Alexander Stark, Federico de Masi, Toby J. Gibson, Joe Lewis, Luis Serrano, Robert B. Russell. “Systematic Discovery of New Recognition Peptides Mediating Protein Interaction Networks,” PLoS Biology 3, 2090-2099 (2005)