Current Topics in Genome Analysis

Lecture 13
December 9, 1997

Protein Sequence Analysis:
Beyond BLAST

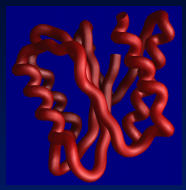
Andy Baxevanis, Ph.D.
andy@nhgri.nih.gov

The Flow of Biotechnology Information

Gene

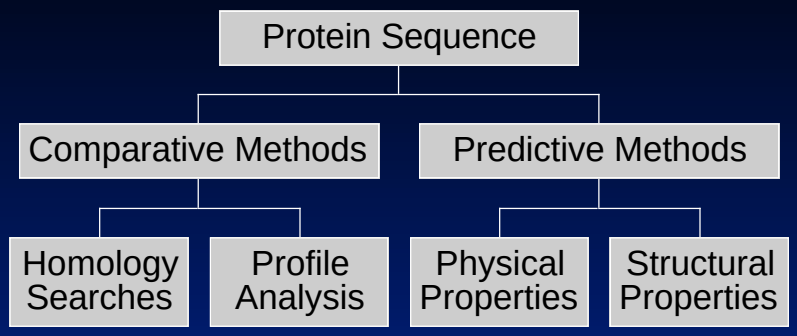
```
> DNA sequence
AATTCATGAAAATCGTATACTGGTCTGGTACCGGCAACAC
TGAGAAAATGGCAGAGCTCATCGCTAAAGGTATCATCGAA
TCTGGTAAAGACGTCAACACCATCAACGTGTCTGACGTTA
ACATCGATGAACTGCTGAACGAAGATATCCTGATCCTGGG
TTGCTCTGCCATGGGCGATGAAGTTCTCGAGGAAAGCGAA
TTTGAACCGTTCATCGAAGAGATCTCTACCAAAATCTCTG
GTAAGAAGGTTGCGCTGTTCGGTTCTTACGGTTGGGGCGA
CGGTAAGTGGATGCGTGACTTCGAAGAACGTATGAACGGC
TACGGTTGCGTTGTTGTTGAGACCCCGCTGATCGTTCAGA
ACGAGCCGGACGAAGCTGAGCAGGACTGCATCGAATTTGG
TAAGAAGATCGCGAACATCTAGTAGA
```

Function



```
> Protein sequence
MKIVYWSGTGNTEKMAELIAKGIIESGKDVNTINVSDVNI
DELLNEDILILGCSAMGDEVLEESEFEPFIEEISTKISGK
KVALFGSYGWGDGKWMRDFEERMNGYGCVVVETPLIVQNE
PDEAEQDCIEFGKKIANI
```

Protein Sequence Analysis



- *Shared ancestry?*
- *Similar function?*
- *Complete sequence?*
- *Beaten to it?*

BLAST

- Seeks high-scoring segment pairs (HSP)
  - pair of sequences that can be aligned without gaps
  - when aligned, have maximal aggregate score (score cannot be improved by extension or trimming)
  - score must be above score threshhold *S*
- Search engines
  - WWW search form
    http://www.ncbi.nlm.nih.gov/BLAST
  - Unix command line
    `blastp nr seqfile [parameters] > outfile`
  - E-mail server
    blast@ncbi.nlm.nih.gov

BLAST Algorithms

Program	*Query Sequence*	*Target Sequence*
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Nucleotide, six-frame translation	Protein
TBLASTN	Protein	Nucleotide, six-frame translation
TBLASTX	Nucleotide, six-frame translation	Nucleotide, six-frame translation

Neighborhood Words

Query Word (W = 3)

↓

```
Query:   GSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAE
```

↓

Neighborhood Words

PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12
etc.	

Neighborhood Score Threshold (T = 13)

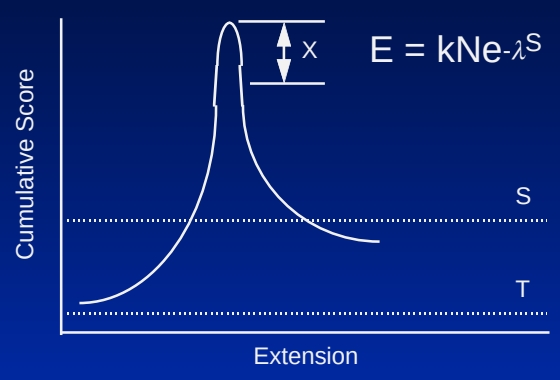
High-Scoring Segment Pairs

PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12
etc.	

```
Query:   325  SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA  365
              +LA++L    TP G R++ +W+  P+ D   + ER   + A
Sbjct:   290  TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA  330
```

BLAST Search Requirements

- A query sequence, in FASTA format
- Which BLAST program to use
- Which database to search
- Parameter values



Scoring Matrices

- Empirical weighting scheme to represent biology
  - Cys/Pro important for structure and function
  - Trp has bulky side chain
  - Lys/Arg have positively-charged side chains
- Importance of understanding scoring matrices
  - Appear in all analyses involving sequence comparison
  - Implicitly represent a particular theory of evolution
  - Choice of matrix can strongly influence outcomes

Matrix Structure

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

PAM Matrices

- Margaret Dayhoff, 1978
- Point Accepted Mutation (PAM)
  - Look at patterns of substitutions in related proteins
  - The new side chain must function the same way as the old one ("acceptance")
  - On average, 1 PAM corresponds to 1 amino acid change per 100 residues
  - 1 PAM ~ 1% divergence
  - Extrapolate to predict patterns at longer distances

PAM Matrices

- Assumptions
  - Replacement is independent of surrounding residues
  - Sequences being compared are of average composition
  - All sites are equally mutable
- Sources of error
  - Small, globular proteins used to derive matrices (departure from average composition)
  - Errors in PAM 1 are magnified up to PAM 250
  - Does not account for conserved blocks or motifs

BLOSUM Matrices

- Henikoff and Henikoff, 1992
- Blocks Substitution Matrix (BLOSUM)
  - Look only for differences in conserved, ungapped regions of a protein family
  - More sensitive to structural or functional substitutions
  - BLOSUM *n*
    - Contribution of sequences > *n*% identical weighted to 1
    - Reduces contribution of closely-related sequences
    - Increasing *n* ~ increasing PAM distance

So many matrices...

- Triple-PAM strategy *(Altschul, 1991)*
  - PAM 40 Short alignments, highly similar
  - PAM 120 ↓
  - PAM 250 Longer, weaker local alignments
- BLOSUM 62 *(Henikoff, 1993)*
  - Most effective in detecting known members of a protein family
  - BLAST default
- No single matrix is the complete answer for all sequence comparisons

BLAST Query

```
>N-terminal unknown protein
MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSASSLLTAAFGNL
FGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQQQHNNN...
```

BLASTP / nr / BLOSUM62

```
                                                                    Smallest
                                                                      Sum
                                                              High  Probability
Sequences producing High-scoring Segment Pairs:              Score  P(N)      N

sp|P29617|PRO_DROME  PROTEIN PROSPERO /pir||S24548 homeot...  1381  1.9e-172  1
pir||JQ1397          pros protein - fruit fly (Drosophila...  1381  1.9e-172  1
pir||A41089          neuronal precursor protein - fruit f...  1110  6.9e-137  1
sp|P29555|HMAA_DROME HOMEOBOX PROTEIN ABDOMINAL-A /gi|969...    80  5.8e-06   3
sp|P54681|RTOA_DICDI RTOA PROTEIN (RATIO-A) /gi|1206019 (...    89  1.4e-05   2
pir||A49070          ecdysone-inducible protein E78A - fr...    89  1.6e-05   4
gnl|PID|e251949      (X98881) nuclear hormone receptor [D...    89  1.6e-05   4
gi|899254            (Z50038) predicted trithorax protein...    92  2.3e-05   2
```

Lower probability infers greater significance – but always look at the alignments!

BLAST Query

```
>N-terminal unknown protein
MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSASSLLTAAFGNL
FGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQQQHNNN...
```

BLASTP / nr / BLOSUM62

```
                                                                    Smallest
                                                                      Sum
                                                              High  Probability
Sequences producing High-scoring Segment Pairs:              Score  P(N)      N

sp|P29617|PRO_DROME  PROTEIN PROSPERO /pir||S24548 homeot...  1381  1.9e-172  1
pir||JQ1397          pros protein - fruit fly (Drosophila...  1381  1.9e-172  1
pir||A41089          neuronal precursor protein - fruit f...  1110  6.9e-137  1
sp|P29555|HMAA_DROME HOMEOBOX PROTEIN ABDOMINAL-A /gi|969...    80  5.8e-06   3
sp|P54681|RTOA_DICDI RTOA PROTEIN (RATIO-A) /gi|1206019 (...    89  1.4e-05   2
pir||A49070          ecdysone-inducible protein E78A - fr...    89  1.6e-05   4
gnl|PID|e251949      (X98881) nuclear hormone receptor [D...    89  1.6e-05   4
gi|899254            (Z50038) predicted trithorax protein...    92  2.3e-05   2
```

Sum of n HSPs:
- *Must be in same orientation*
- *Must not overlap*
- *Often seen with repeated motifs*

Profiles

- Numerical representations of multiple sequence alignments
- Depend upon *patterns* or *motifs* containing conserved residues
- Represent the common characteristics of a protein family
- Can find similarities between sequences with little or no sequence identity
- Allow for the analysis of distantly-related proteins

Profile Construction

```
APHIIVATPG
GCEIVIATPG
GVEICIATPG
GVDILIGTTG
RPHIIVATPG
KPHIIIATPG
KVQLIIATPG
RPDIVIATPG
APHIIVGTPG
APHIIVGTPG
GCHVVIATPG
NQDIVVATTG
```

- *Which residues are seen at each position?*
- *What is the frequency of observed residues?*
- *Which positions are conserved?*
- *Where can gaps be introduced?*

Position-Specific Scoring Table

Cons	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z
G	17	18	0	19	14	-22	31	0	-9	12	-15	-5	15	10	9	6	18	14	1	-15	-22	11
P	18	0	13	0	0	-12	13	0	8	-3	-3	-1	-2	23	2	-2	12	11	17	-31	-8	1
H	5	24	-12	29	25	-20	8	32	-9	9	-10	-9	22	7	30	10	0	4	-8	-20	-7	27
I	-1	-12	6	-13	-11	33	-12	-13	63	-11	40	29	-15	-9	-14	-15	-6	7	50	-17	8	-11
V	3	-11	1	-11	-9	22	-3	-11	46	-9	37	30	-13	-3	-9	-13	-6	6	50	-19	2	-8
V	5	-9	9	-9	-9	19	-1	-13	57	-9	35	26	-13	-2	-11	-13	-4	9	58	-29	0	-9
A	54	15	12	20	17	-24	44	-6	-4	-1	-11	-5	12	19	9	-13	21	19	9	-39	-20	10
T	40	20	20	20	20	-30	40	-10	20	20	-10	0	20	30	-10	-10	30	150	20	-60	-30	10
P	31	6	7	6	6	-41	19	11	-9	6	-16	-11	0	89	17	17	24	22	9	-50	-48	12
G	70	60	20	70	50	-60	150	-20	-30	-10	-50	-30	40	30	20	-30	60	40	20	-100	-70	30

ProfileScan

- Search sequence against a collection of profiles
- Databases available
  - PROSITE 1167 entries
  - Pfam 527 entries
- *http://www.ch.embnet.org/software/PFSCAN_form.html*

ProfileScan Query

```
>C-terminal end
MALLQISEPGLSAAPHQRRLAAGIDLGTTNSLVATVRSGQAETLADHEGRHLLPSVVHYQQQGHSVGYDA
RTNAALDTANTISSVKRLMGRSLADIQQRYPHLPYQFQASENGLPMIETAAGLLNPVRVSADILKALAAR
ATEALAGELDGVVITVPAYFDDAQRQGTKDAARLAGLHVLRLLNEPTAAAIAYGLDSGQEGVIAVYDLGG
GTFDISILRLSRGVFEVLATGGDSALGGDDFDHLLADYIREQAGIPDRSDNRVQRELLDAAIAAKIA...
```

Prosite + Pfam
Significant matches only

```
normalized raw        from - to  Profile|Description
219.3535 27400 pos.     21 - 600 PF00012|HSP70 Heat shock hsp70 proteins
```

E-value

NScore	SwissProt
7.0	1.8000
8.0	0.1800
9.0	0.0180
10.0	0.0018
219.4	3e-211

Signatures

```
[IV]-D-L-G-T-[ST]-x-[SC]

[LIVMF]-[LIVMFY]-[DN]-[LIVMFS]-G-[GSH]-[GS]-[AST]-x(3)-
[ST]-[LIVM]-[LIVMFC]

[LIVM]-x-[LIVMF]-x-G-G-x-[ST]-x-[LIVM]-P-x-[LIVM]-x-
[DEQKRSTA]
```

BLOCKS

- Steve Henikoff, Fred Hutchinson Cancer Research Center, Seattle
- Multiple alignments of conserved regions in protein families
  - 1 "block" = 1 short, ungapped multiple alignment
  - Families can be defined by one or more blocks
  - Searches allow detection of one or more blocks representing a family
- Search engines
  - E-Mail *blocks@howard.fchrc.org*
  - Web *http://blocks.fhcrc.org/*

BLOCKS Query

```
>C-terminal end
MALLQISEPGLSAAPHQRRLAAGIDLGTTNSLVATVRSGQAETLADHEGRHLLPSVVHYQQQGHSVGYDA
RTNAALDTANTISSVKRLMGRSLADIQQRYPHLPYQFQASENGLPMIETAAGLLNPVRVSADILKALAAR
ATEALAGELDGVVITVPAYFDDAQRQGTKDAARLAGLHVLRLLNEPTAAAIAYGLDSGQEGVIAVYDLGG
GTFDISILRLSRGVFEVLATGGDSALGGDDFDHLLADYIREQAGIPDRSDNRVQRELLDAAIAAKIA...
```

↓ *Search blocks*

```
BL00297A
HSCA_ECOLI 136     ALAARATEALAGELDGVVITVPAYFDDAQRQGTKDAARLAGLHVLRLLNEPTAAA
                   ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
C-terminal 136     ALAARATEALAGELDGVVITVPAYFDDAQRQGTKDAARLAGLHVLRLLNEPTAAA
```

↓ *Examine blocks*

```
ID   HSP70_1; BLOCK
AC   BL00297A; distance from previous block=(94,187)
DE   Heat shock hsp70 proteins family proteins.
BL   PRR motif; width=55; seqs=111; 99.5%=2947; strength=1607
```

BLOCKS Entry

```
ID   HSP70_1; BLOCK
AC   BL00297A; distance from previous block=(94,187)
DE   Heat shock hsp70 proteins family proteins.
BL   PRR motif; width=55; seqs=111; 99.5%=2947; strength=1607
HS70_CHLRE ( 129) KETAQASLGADREVKKAVVTVPAYFNDSQRQATKDAGMIAGLEVLRIINEPTAAA  19

 HS7L_SBYV ( 132) ALISTASEAFKCQCTGVICSVPANYNCLQRSFTESCVNLSGYPCVYMVNEPSAAA  75

HS7R_HUMAN ( 124) KLKETAESVLKKPVVDCVVSVPCFYTDAERRSVMDATQIAGLNCLRLMNETTAVA  45

HS7T_MOUSE ( 126) TKMKETAEVFWAPMSQRVITVPAYFNDSQRQATKDAGVIAGLNVLRIINEPTAVA  28

YKH3_YEAST ( 160) SLLKDRDARTEDFVNKMSFTIPDFFDQHQRKALLDASSITTGIEETYLVSEGMSV 100

DNAK_BACSU (  95) HLKSYAESYLGETVSKAVITVPAYFNDAERQATKDAGKIAGLEVERIINEPTAAA   7
DNAK_BORBU ( 122) KMKETAEAYLGEKVTEAVITVPAYFNDAQRQATKDAGKIAGLEVKRIVNEPTAAA   3
DNAK_BRUOV ( 122) KMKETAESYLGETVTQAVITVPAYFNDAQRQATKDAGKIAGLEVLRIINEPTAAA   3
DNAK_BURCE ( 123) KMKKTAEDYLGEPVTEAVITVPAYFNDSQRQATKDAGRIAGLEVKRIINEPTAAA   3
DNAK_CAUCR ( 122) KMKEAAEAHLGEPVTKAVITVPAYFNDAQRQATKDAGKIAGLEVLRIINEPTAAA   5
DNAK_CHLPN ( 125) KMKETAEAYLGETVTEAVITVPAYFNDSQRASTKDAGRIAGLDVKRIIPEPTAAA  10
DNAK_CLOPE (  98) KLKADAEAYLGEKVTEAVITVPAYFNDAERQATKDAGRIAGLDVKTIINEPTAAS   8
DNAK_CRYPH ( 122) KLVDDASKYLGESVKQAVITVPAYFNDSQRQATKDAGRIAGLEVLRIINEPTAAS   5
DNAK_ECOLI ( 121) KMKKTAEDYLGEPVTEAVITVPAYFNDAQRQATKDAGRIAGLEVKRIINEPTAAA   3
DNAK_ERYRH (  96) YMKSYAEDYLGEKVTKAVITVPAYFNDAQRQATKDAGKIAGLEVERIINEPTAAA   5
DNAK_HAEIN ( 120) KMKKTAEDFLGESVTEAVITVPAYFNDAQRQATIDAGKIAGLDVKRIINEPTAAA   6
                                                .
                                                .
                                                .
```

BLOCK Maker

```
>chk-H5
SRRSASHPTYSEMIAAAIRAEKSRGGSSRQSIQKYIKSHYKVGHNADLQIKLSIRRLLAAGVLKQTKGVGASGSFRLAKS
>hum-H1
TPRKASGPPVSELITKAVAASKERSGVSLAALKKALAAAGYDVEKNNSRIKLGLKSLVSKGTLVQTKGTGASGSFKLNKK
>pea-H1
PRNPASHPTYEEMIKDAIVSLKEKNGSSQYAIAKFIEEKQKQLPANFKKLLLQNLKKNVASGKLIKVKGSFKLSAAAKKP
```

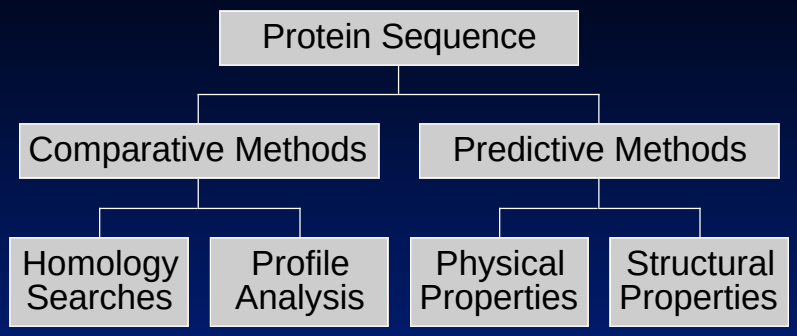
↓ *MOTIF/GIBBS*

```
>Histone chk-H5  family
6 sequences are included in 2 blocks

            HistoneA, width = 31
    chk-H5     1 SHPTYSEMIAAAIRAEKSRGGSSRQSIQKYI
    hum-H1     1 SGPPVSELITKAVAASKERSGVSLAALKKAL
    pea-H1     1 SHPTYEEMIKDAIVSLKEKNGSSQYAIAKFI
  sce-H1.1     1 SSKSYRELIIEGLTALKERKGSSRPALKKFI
  sce-H1.2     1 SSLTYKEMILKSMPQLNDGKGSSRIVLKKYV
    xla-H1     1 SGPSASELIVKAVSSSKERSGVSLAALKKAL

            HistoneB, width = 15
    chk-H5  (  21)    53 IRRLLAAGVLKQTKG
    hum-H1  (  21)    53 LKSLVSKGTLVQTKG
    pea-H1  (  21)    53 LKKNVASGKLIKVKG
  sce-H1.1  (  21)    53 IKKGVEAGDFEQPKG
  sce-H1.2  (  21)    53 IKKCVENGELVQPKG
    xla-H1  (  21)    53 LKALVTKGTLTQVKG
```

Protein Sequence Analysis

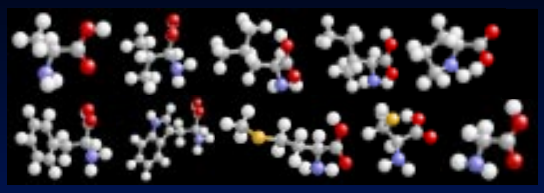


- *Composition*
- *Hydrophobicity*
- *Secondary structure*
- *Specialized structures*
- *Tertiary structure*

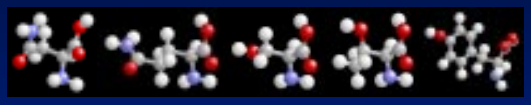
Information Landscape

Information Landscape

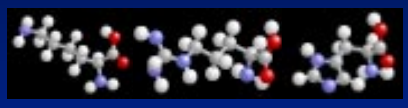
Nonpolar



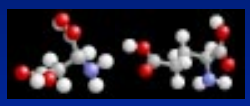
Polar *Neutral*



Polar *Basic*



Polar *Acidic*



ProtParam

- Computes physicochemical parameters
  - Molecular weight
  - Theoretical pI
  - Amino acid composition
  - Extinction coefficient
- Simple query
  - SWISS-PROT accession number
  - User-entered sequence, in single-letter format
- *http://expasy.hcuge.ch/sprot/protparam.html*

ProtParam Query

```
MNGEADCPTDLEMAAPKGQDRWSQEDMLTLLECMKNNLPSNDSSKFKTTESHMDWEKVAFKDFSGDMCKL
KWVEISNEVRKFRTLTELILDAQEHVKNPYKGKKLKKHPDFPKKPLTPYFRFFMEKRAKYAKLHPEM...
```

↓ *Compute parameters*

```
Number of amino acids: 727
Molecular weight: 84936.8
Theoretical pI: 5.44

Amino acid composition:

Ala (A)  35       4.8%        Leu (L)  57       7.8%
Arg (R)  39       5.4%        Lys (K)  97      13.3%
Asn (N)  28       3.9%        Met (M)  25       3.4%
Asp (D)  58       8.0%        Phe (F)  18       2.5%
Cys (C)   6       0.8%        Pro (P)  39       5.4%
Gln (Q)  36       5.0%        Ser (S)  67       9.2%
Glu (E)  98      13.5%        Thr (T)  22       3.0%
Gly (G)  26       3.6%        Trp (W)  11       1.5%
His (H)  11       1.5%        Tyr (Y)  20       2.8%
Ile (I)  18       2.5%        Val (V)  16       2.2%

Asx (B)   0       0.0%
Glx (Z)   0       0.0%
Xaa (X)   0       0.0%

Total number of negatively charged residues (Asp + Glu): 156
Total number of positively charged residues (Arg + Lys): 136
```

PROPSEARCH

- Uses amino acid composition to detect weak relationships
- Can be used to discern members of the same protein family
- 144 physical properties used in analysis (“vector”)
  - Molecular weight
  - Bulky residue content
  - Average hydrophobicity and charge
- Search against “database of vectors” (PIR and SWISS-PROT)
- *http://www.embl-heidelberg.de/prs.html*

PROPSEARCH Query

```
>S18193 autoantigen NOR-90 - human
MNGEADCPTDLEMAAPKGQDRWSQEDMLTLLECMKNNLPSNDSSKFKTTESHMDWEKVAFKDFSGDMCKL
KWVEISNEVRKFRTLTELILDAQEHVKNPYKGKKLKKHPDFPKKPLTPYFRFFMEKRAKYAKLHPEM...
```

Vector search

DIST	Odds
< 10	87.0%
< 8.7	94.0%
< 7.5	99.6%

		DIST	LEN2	POS1	POS2	pI	DE
		0.00	727	1	727	5.33	autoantigen NOR-90 - human
		1.36	764	1	764	5.62	NUCLEOLAR TRANSCRIPTION FACTOR 1
		1.40	765	1	765	5.55	NUCLEOLAR TRANSCRIPTION FACTOR 1
		1.57	764	1	764	5.61	NUCLEOLAR TRANSCRIPTION FACTOR 1
		3.95	677	1	677	5.79	NUCLEOLAR TRANSCRIPTION FACTOR 1
6	[illegible]	4.18	701	1	701	6.05	NUCLEOLAR TRANSCRIPTION FACTOR 2
7	>p1;s57[illegible]	7.72	606	1	606	6.63	hypothetical protein YPR018w - yeast
8	>p1;i50463	8.49	772	1	772	5.71	protein kinase - chicken
9	>p1;h54024	8.83	768	1	768	5.27	protein kinase (EC 2.7.1.37) cdc2-related
10	>p1;b54024	8.87	777	1	777	5.27	protein kinase (EC 2.7.1.37) cdc2-related
11	>p1;g54024	8.90	766	1	766	5.21	protein kinase (EC 2.7.1.37) cdc2-related
12	>p1;a55817	9.00	783	1	783	5.19	cyclin-dependent kinase p130-PITSLRE - mouse
13	>p1;f54024	9.11	777	1	777	5.30	protein kinase (EC 2.7.1.37) cdc2-related
14	>p1;e54024	9.11	779	1	779	5.42	protein kinase (EC 2.7.1.37) cdc2-related
15	yaa5_schpo	9.45	598	1	598	4.78	HYPOTHETICAL 69.5 KD PROTEIN C22G7.05
16	>p1;s62449	9.45	598	1	598	4.78	hypothetical protein SPAC22G7.05 - fission yeast
17	>f1;i58390	9.45	920	1	920	5.00	retinoblastoma binding protein 1 isoform I
18	>p1;s63193	9.58	590	1	590	6.15	hypothetical protein YNL227c - yeast
19	ynw7_yeast	9.58	590	1	590	6.15	HYPOTHETICAL 68.8 KD PROTEIN IN URE2-SSU72
20	>p1;s49634	9.74	899	1	899	4.79	hypothetical protein YML093w - yeast
21	ymj3_yeast	9.74	899	1	899	4.79	HYPOTHETICAL 103.0 KD PROTEIN IN RAD10-PRS4
22	radi_human	9.76	583	1	583	6.33	RADIXIN.
23	radi_pig	9.81	583	1	583	6.21	RADIXIN (MOESIN B).
24	>f1;i78883	9.83	866	1	866	4.77	retinoblastoma binding protein 1 isoform II
25	>p1;b42997	9.87	754	1	754	5.17	retinoblastoma-associated protein 2 - human
26	>p1;a57467	9.91	647	1	647	5.74	RalBP1 - rat

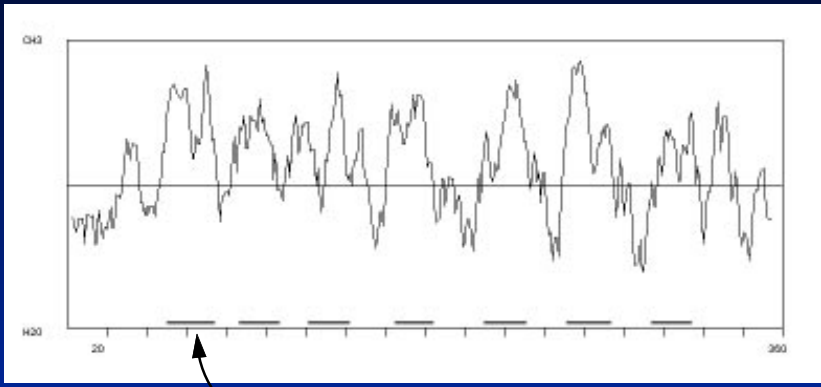
TGREASE

- Calculates hydrophobicity along length of a protein *(Kyte and Doolittle, 1982)*
- Hydropathy scale
  - Propensity to bury side chain within protein core
  - Based on solubility, free energy of transfer through water-vapor transition, and other factors
  - More positive scores indicate greater hydrophobicity
  - More negative scores indicate greater hydrophilicity
- Moving average, with 7-11 residues optimal
- *ftp://ftp.virginia.edu/pub/fasta*

TGREASE Output

```
>sp|P25025|IL8B_HUMAN HIGH AFFINITY INTERLEUKIN-8 RECEPTOR B
MEDFNMESDSFEDFWKGEDLSNYSYSSTLPPFLLDAAPCEPESLEINKYFVVIIYALVFLLSLLGNSLVM
LVILYSRVGRSVTDVYLLNLALADLLFALTLPIWAASKVNGWIFGTFLCKVVSLLKEVNFYSGILLL
```

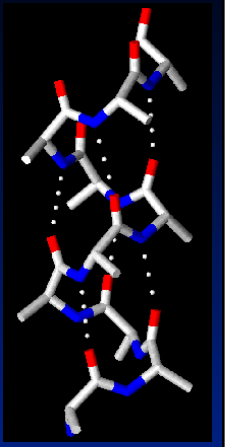
Default window length (11)



Transmembrane segments from SWISS-PROT

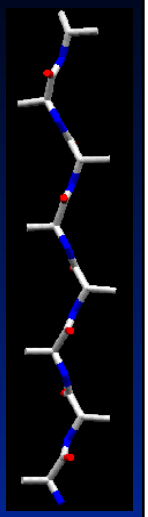
Alpha-helix

- Corkscrew
- Main chain forms backbone, side chains project out
- Hydrogen bonds between CO group at *n* and NH group at *n+4*
- Helix-formers: Ala, Glu, Leu, Met
- Helix-breaker: Pro



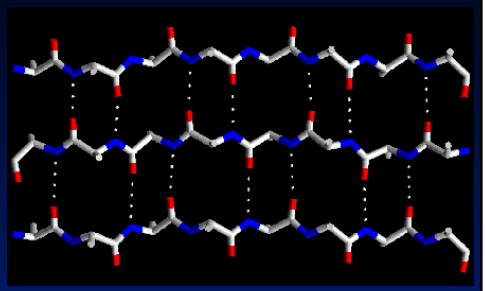
Beta-strand

- Extended structure ("pleated")
- Peptide bonds point in opposite directions
- Side chains point in opposite directions
- No hydrogen bonding *within* strand



Beta-sheet

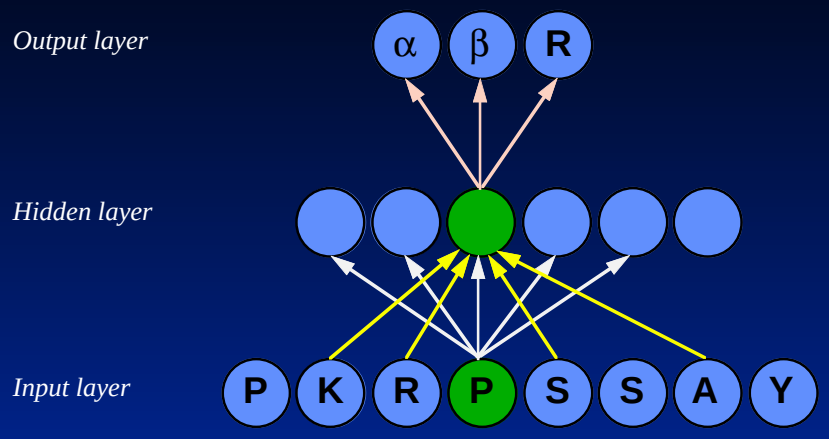
- Stabilization through hydrogen bonding
- Parallel or antiparallel
- Variant: beta-turn



Folding Classes

α	β	α+β	α/β
2CCY	1CID	1SNC	1TIM
Cyt c	*CD4*	*Staph nuclease*	*Triose phosphate isomerase*

Neural Network



nnpredict

- Neural network approach to making predictions *(Kneller et al., 1990)*
- Best-case accuracy > 65%
- Search engines
  - E-mail *nnpredict@celeste.ucsf.edu*
  - Web *http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html*

nnpredict Query

```
option: a/b
>flavodoxin - Anacystis nidulans
AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADASDLNAYDYLIIGCPTWNVGELQSDWEGIY
DDLDSVNFQGKKVAYFGAGDQVGYSDNFQDAMGILEEKISSLGSQTVGYWPIEGYDFNESKAVRNNQFVG
LAIDEDNQPDLTKNRIKTWVSQLKSEFGL
```

↓ *α/β folding class*

```
Tertiary structure class: alpha/beta

Sequence:
AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADASDLNAYDYLIIGCPTWNVG
ELQSDWEGIYDDLDSVNFQGKKVAYFGAGDQVGYSDNFQDAMGILEEKISSLGSQTVGYW
PIEGYDFNESKAVRNNQFVGLAIDEDNQPDLTKNRIKTWVSQLKSEFGL

Secondary structure prediction (H = helix, E = strand, - = no prediction):
----EEE------EEEHHHHHHH------EEEH---------------EEEE--------
---------------HHHH---EEEE------------H--HHHHHHHH------E--E-
-E--------------HH--E---------------EHHHHH------
```

PredictProtein

- Multi-step predictive algorithm *(Rost et al., 1994)*
  - Protein sequence queried against SWISS-PROT
  - MaxHom used to generate iterative, profile-based multiple sequence alignment *(Sander and Schneider, 1991)*
  - Multiple alignment fed into neural network (PHDsec)
- Accuracy
  - Average > 70%
  - Best-case > 90%
- Search engines
  - E-mail *predictprotein@embl-heidelberg.de*
  - Web *http://www.embl-heidelberg.de/predictprotein/*

PredictProtein Query

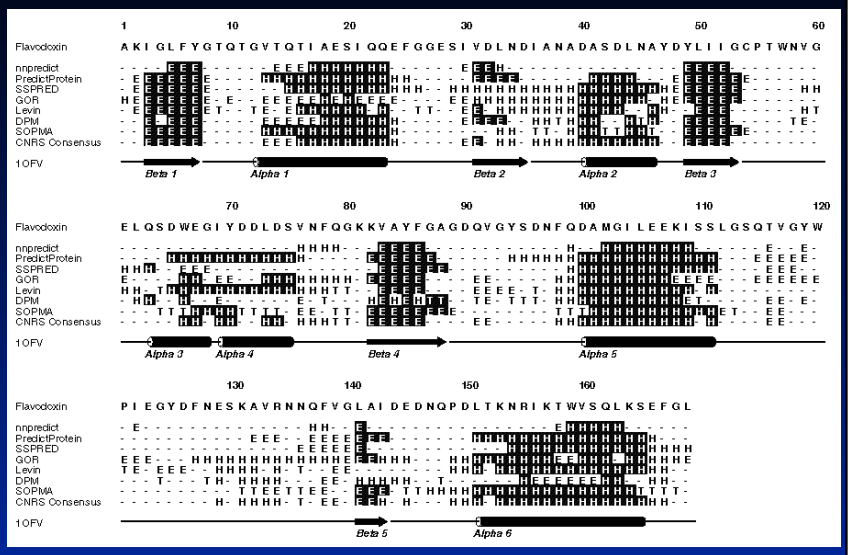
```
Joe Buzzcut
National Human Genome Research Institute, NIH
buzzcut@nhgri.nih.gov
# flavodoxin - Anacystis nidulans
AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADASDLNAYDYLIIGCPTWNVGELQSDWEGIY
DDLDSVNFQGKKVAYFGAGDQVGYSDNFQDAMGILEEKISSLGSQTVGYWPIEGYDFNESKAVRNNQFVG
LAIDEDNQPDLTKNRIKTWVSQLKSEFGL
```

Secondary structure

```
          ....,....1....,....2....,....3....,....4....,....5....,....6
AA       |AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADASDLNAYDYLIIGCPTWNVG|
PHD sec  | EEEEEEE    HHHHHHHHHHHHH     EEEEE HHH HHHH    EEEEE       |
Rel sec  |9389997369824899999999997679824432132412786312419998615477 65|
Detail:
prH sec  |0000000000146899999999998210000111125653887643210000011111 11|
prE sec  |0589988520000000000000000000003665542100000000001489987412 0002|
prL sec  |9310001379853100000000001789852223443245112345540001146677 76|
```

- *SWISS-PROT hits*
- *Multiple alignment*
- *PDB homologues*

Accuracy of Predictions



SignalP

- Neural network trained based on phylogeny
  - Gram-negative prokaryotic
  - Gram-positive prokaryotic
  - Eukaryotic
- Predicts secretory signal peptides (*not* those involved in intracellular signal transduction)
- *http://www.cbs.dtu.dk/services/SignalP*

SignalP Query

```
>sp|P05019|IGFB_HUMAN INSULIN-LIKE GROWTH FACTOR IB PRECURSOR
MGKISSLPTQLFKCCFCDFLKVKMHTMSSSHLFYLALCLLTFTSSATAGPETLCGAELVDALQFVCGDRG
```

N-terminal end only
Eukaryotic set

```
************************ SignalP predictions ************************
Using networks trained on euk data

>IGF-IB        length = 195

# pos  aa    C       S       Y
      .
      .
      .
   46   A   0.365   0.823   0.495
   47   T   0.450   0.654   0.577
   48   A   0.176   0.564   0.369
   49   G   0.925   0.205   0.855
   50   P   0.185   0.163   0.376
      .
      .
      .
< Is the sequence a signal peptide?
# Measure  Position  Value  Cutoff  Conclusion
  max. C    49       0.925   0.37   YES
  max. Y    49       0.855   0.34   YES
  max. S    37       0.973   0.88   YES
  mean S     1-48    0.550   0.48   YES
# Most likely cleavage site between pos. 48 and 49: ATA-GP
```

C = cleavage site score
S = signal peptide score
Y = combined score

PHDtopology

- Approach similar to PredictProtein (PHDsec)
- Overall two-state accuracy 94.7%
  - Accuracy of predicting helix 92.0%
  - Accuracy of predicting loop 96.0%
- Includes topology prediction
- Search engines
  - E-mail *predictprotein@embl-heidelberg.de*
  - Web *http://www.embl-heidelberg.de/predictprotein/*

PHDtopology Query

```
Joe Buzzcut
National Human Genome Research Institute, NIH
buzzcut@nhgri.nih.gov
predict htm topology
# pendrin
MAAPGGRSEPPQLPEYSCSYMVSRPVYSELAFQQQHERRLQERKTLRESLAKCCSCSRKRAFGVLKTLVPILEWLPKYRV
KEWLLSDVISGVSTGLVATLQGMAYALLAAVPVGYGLYSAFFPILTYFIFGTSRHISVGPFPVVSLMVGSVVLSMAP...
```

```
                 ....,....37...,....38...,....39...,....40...,....41...,....42
         AA     |YSLKYDYPLDGNQELIALGLGNIVCGVFRGFAGSTALSRSAVQESTGGKTQIAGLIGAII|
         PHD htm |                      HHHHHHHHHHHHHH              HHHHHHHHHH|
         Rel htm |36889999999999999864110466777765543125777888777621467788888|
 detail:         |                                                            |
         prH htm |31000000000000000012445788888887776532111000011135788899999|
         prL htm |68999999999999999987554211111112223467888999988864211100000|
 subset:         |                                                            |
         SUB htm |.LLLLLLLLLLLLLLLLL.....HHHHHHHHH....LLLLLLLLLLL...HHHHHHHH|
  other:         |                                                            |
         PHDFhtm |                      HHHHHHHHHHHHHH              HHHHHHHHHH|
         PHDRhtm |                   HHHHHHHHHHHHHHHHHH               HHHHHHHH|
         PHDThtm |iiiiiiiiiiiiiiiiiiTTTTTTTTTTTTTTTTTTooooooooooooooooTTTTTTTT|
```

Predicting Tertiary Structure

- Sequence specifies conformation, *but* conformation does *not* specify sequence
- Structure is conserved to a much greater extent than sequence
- Similarities between proteins may not necessarily be detected through “traditional” methods
- Protein folding problem
  - Asilomar structure prediction “contest”
  - Numerous protein folds can be reliably identified
  - Consensus approach

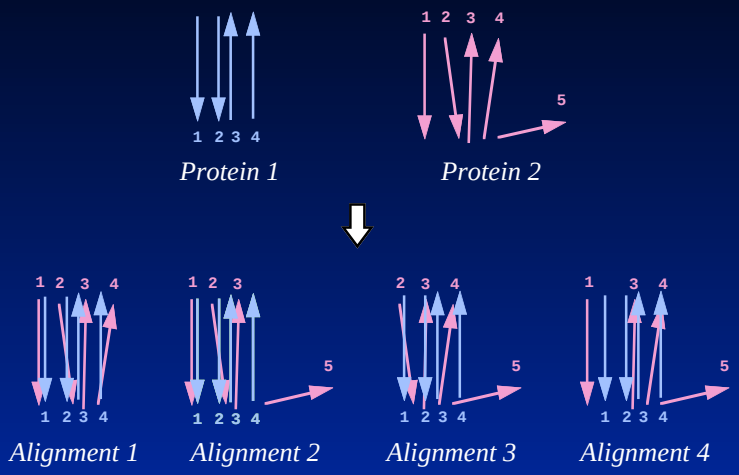
VAST Structure Comparison

Step 1: Construct vectors for secondary structure elements

Ricin Chain B

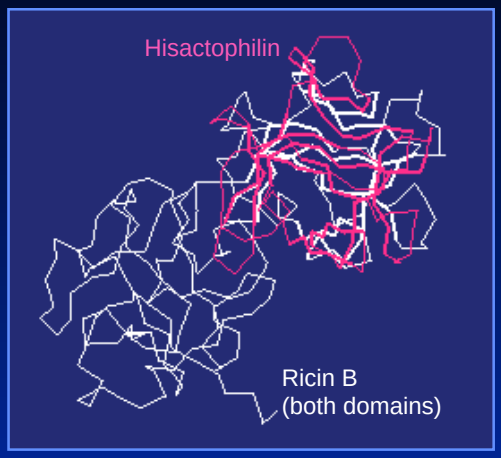
VAST Structure Comparison

Step 2: Optimally align structure element vectors

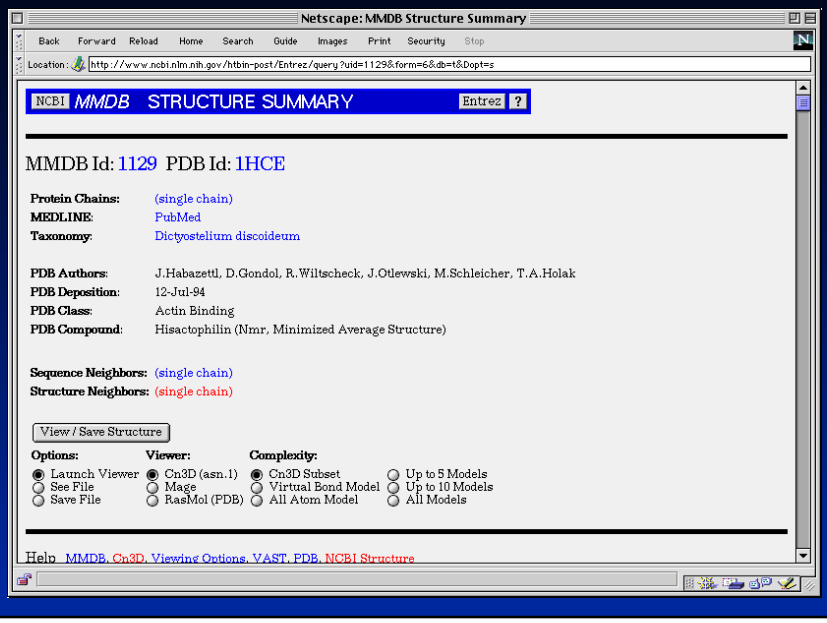


VAST Structure Comparison

Step 3: Refine residue-by-residue alignment using Monte Carlo



VAST Query



VAST Results

Netscape: Vast Results

Back Forward Reload Home Search Guide Images Print Security Stop

Location: http://www.ncbi.nlm.nih.gov/cgi-bin/Structure/vastsrv?uid=11298&chaindom=11290100

NCBI VAST

VAST Homepage and table legends.

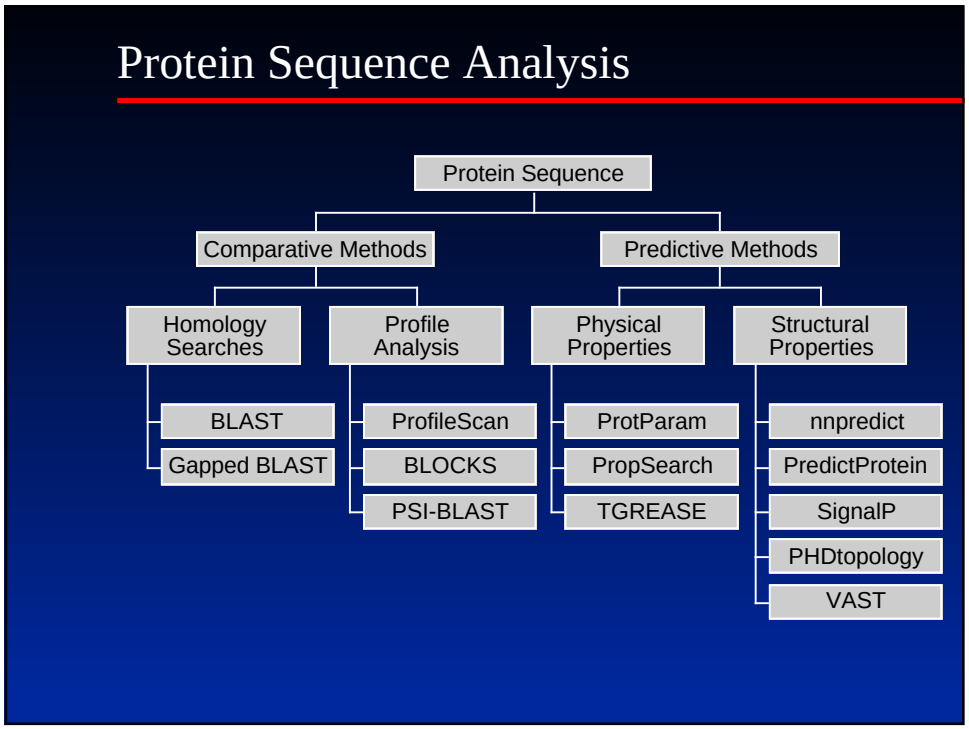
Similar structures to 1HCE

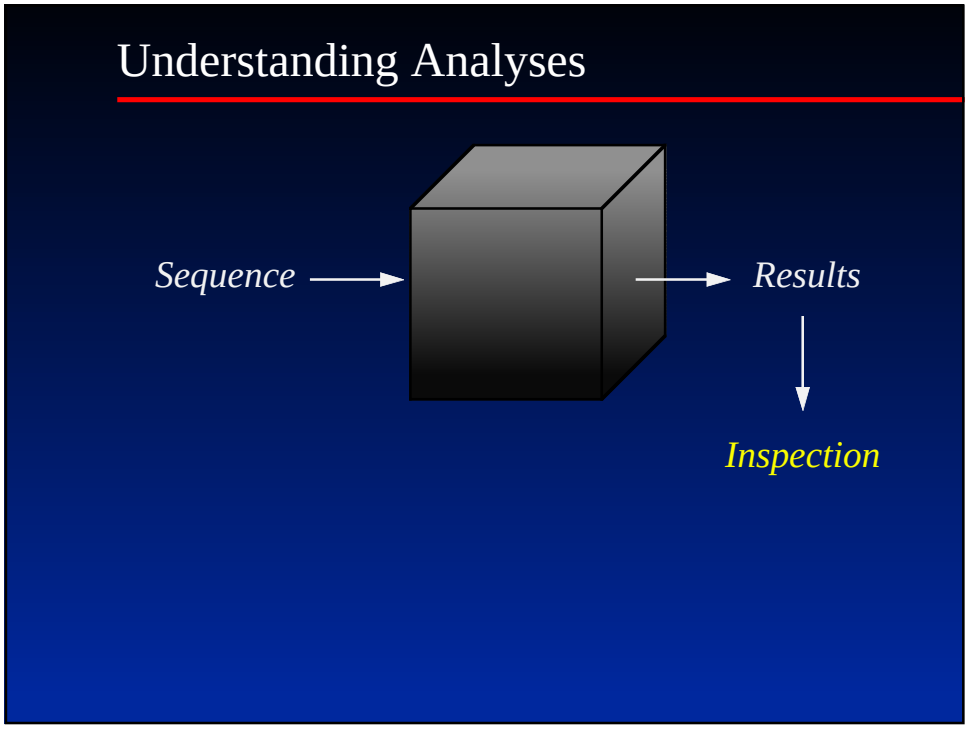
Neighbors of: 1HCE
Hisactophilin (Nmr, Minimized Average Structure)

Structure	C	D	A	SCO	PVAL	RMSD	NRES	Id	Contents
(P) (K) 1HCD			1	20.7	12.3	1.4	118	100.0	Hisactophilin (Nmr, 1 Structure)
(P) (K) 1ABR	B	3	1	18.4	10.4	2.2	104	7.7	Abrin-A Complexed With Two Sugar Chains
(P) (K) 1ABR	B		1	18.4	7.9	2.2	104	7.7	Abrin-A Complexed With Two Sugar Chains
(P) (K) 8I1B			1	18.2	10.1	2.6	111	17.1	Interleukin 1-Beta
(P) (K) 1AFC	C		1	17.8	9.9	2.2	101	13.9	Acidic Fibroblast Growth Factor (Afgf) Mutant With Cys 47 Replaced By Ala (C47a) Complex With Sucrose Octasulfate
(P) (K) 2AAI	B	1	1	17.8	9.9	2.2	104	8.6	Ricin (E.C.3.2.2.22)
(P) (K) 1AFC	F		1	17.8	9.9	2.3	106	13.2	Acidic Fibroblast Growth Factor (Afgf) Mutant With Cys 47 Replaced By Ala (C47a) Complex With Sucrose Octasulfate
(P) (K) 2AAI	B		1	17.8	7.3	2.2	104	8.6	Ricin (E.C.3.2.2.22)
(P) (K) 5I1B			1	17.7	9.9	2.6	113	18.6	Interleukin 1-Beta
(P) (K) 1ILT	B		1	17.7	9.9	2.9	113	12.4	Interleukin-1 Receptor Antagonist (Il-1ra) (Alpha Carbons)
(P) (K) 2IRT	A		1	17.7	9.9	2.5	109	11.9	Interleukin-1 Receptor Antagonist Protein
(P) (K) 1AFC	E		1	17.6	8.9	2.2	104	13.5	Acidic Fibroblast Growth Factor (Afgf) Mutant With Cys 47 Replaced By Ala (C47a) Complex With Sucrose Octasulfate
(P) (K) 41BI			1	17.4	9.7	2.7	113	18.6	Interleukin-1Beta (IL-1Beta) (Mutant With Cys 8 Replaced By Ala (C8A)

Tertiary Structure Prediction

- Homology model building
  (Bryant and Lawrence, 1993; Jones and Thornton, 1996)
- SWISS-MODEL
  http://expasy.hcuge.ch/swissmod/SWISS-MODEL.html
- DALI
  http://www.embl-heidelberg.de/dali/dali.html
- TOPITS
  http://www.embl-heidelberg.de/predictprotein/phd_help.html


Protein Sequence Analysis
Protein Sequence
Comparative Methods
Predictive Methods
Homology Searches
Profile Analysis
Physical Properties
Structural Properties
BLAST
Gapped BLAST
ProfileScan
BLOCKS
PSI-BLAST
ProtParam
PropSearch
TGREASE
nnpredict
PredictProtein
SignalP
PHDtopology
VAST


Understanding Analyses
Sequence
Results
Inspection