Current Topics in Genome Analysis
Fall 2003

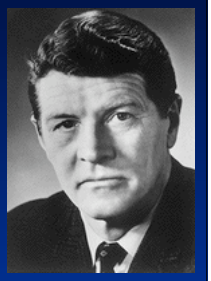
Week 5
Biological Sequence Analysis II

Andy Baxevanis, Ph.D.

Overview

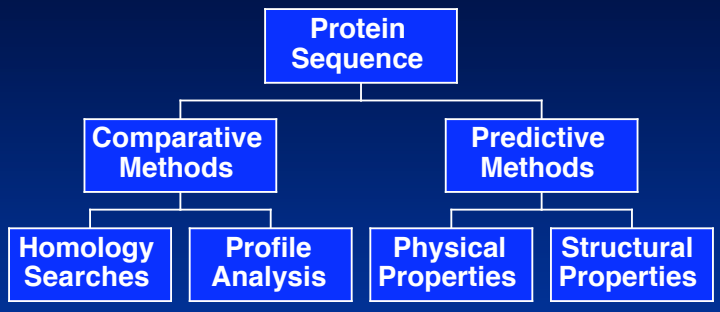
- Week 4: Comparative methods and concepts
  - Similarity *vs*. Homology
  - Global *vs*. Local Alignments
  - Dotplots
  - Scoring Matrices
  - BLAST
- Week 5: Predictive methods and concepts
  - Profiles, patterns, motifs, and domains
  - Secondary structure prediction
  - Structures: VAST, Cn3D, and *de novo* prediction

Protein Conformation



- Christian Anfinsen
  Studies on reversible denaturation →
  "Sequence specifies conformation"
- Chaperones and disulfide interchange enzymes:
  involved but not controlling final state
- "Starting with a newly-determined sequence, what can be determined computationally about its possible function and structure?"

Protein Sequence Analysis



- *Common structure?*
- *Common function?*
- *Evolutionary relationship?*
- *Global or local similarity?*

Sequence Comparisons

- Homology searches
  - Usually "one-against-one" — BLAST, FASTA
  - Allows for comparison of individual sequences against databases comprised of individual sequences
- Profile searches
  - Uses collective characteristics of a family of proteins
  - Search can be "one-against-many" — ProfileScan, CDD

    or "many-against-one" — PSI-BLAST

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	13	6	15	6	6	-12	13	6	6	-5	-5	-1	6	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	11	19	11	2	6	16	11	6	89	17	17	24	22	9	-50	-48	12
G	73	63	23	73	53	-8	150	-20	-30	-10	-50	-30	40	30	20	-30	60	40	20	-100	-70	30

Patterns

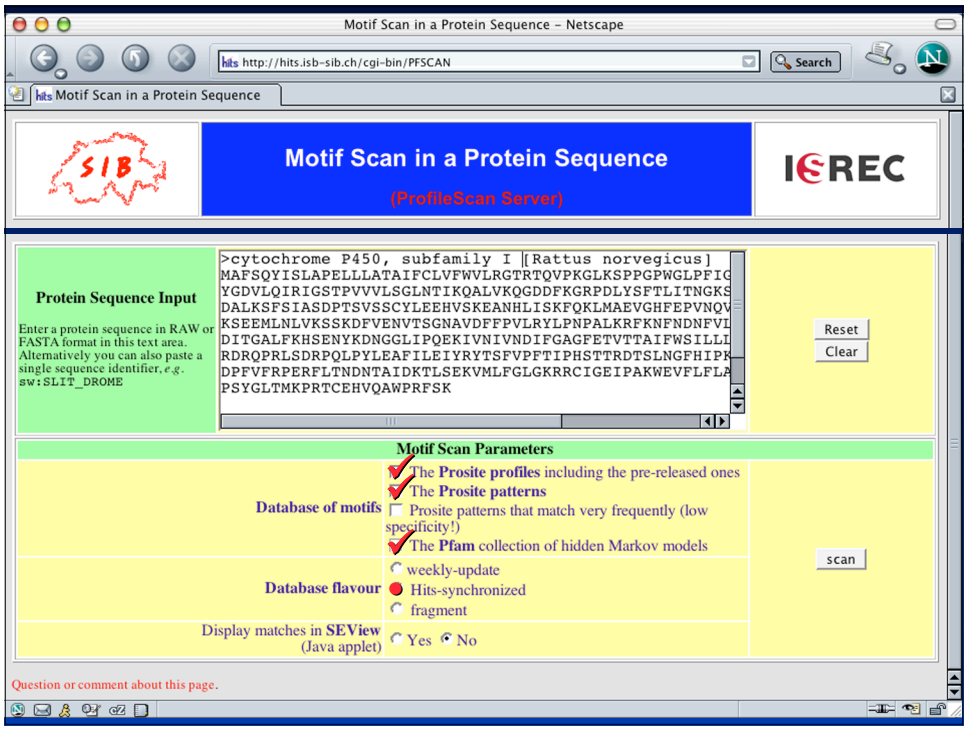
`[FY]-x-C-x(2)-[VA]-x-H(3)`

reads as:

Phe *or* Tyr	**followed by**
any amino acid	**followed by**
Cys	**followed by**
any two amino acids	**followed by**
Val *or* Ala	**followed by**
any amino acid	**followed by**
three His	

ProfileScan

- Search sequence against a collection of profiles and patterns
- Databases available
  - PROSITE profiles
  - PROSITE patterns
  - PfamA
  - PfamB
- *http://hits.isb-sib.ch/cgi-bin/PFSCAN*



Motif Scan in a Protein Sequence – Netscape

http://hits.isb-sib.ch/cgi-bin/PFSCAN_parser

Motif Scan in a Protein Sequence

- **Databases:** Prosite patterns (Hits-synchronized), Prosite profiles (Hits-synchronized), Pfam collection of hidden Markov models (Hits-synchronized)

...scanning for hits_pattern
...scanning for hits_profile
...scanning for hits_pfam
...confirming pfam

Result

- **Summary**:

? pat:CYTOCHROME_P450 pos. **449 - 458**
! pfam:P450 pos. **41 - 506** E-value=**4.2e-138**

status: !
pos.: **41-506**
raw-score = **472.2**
N-score = **144.703**
E-value = **4.2e-138**

pfam:P450
Cytochrome P450
PF00067
Pfam-site ❶
InterPro. ❷

- **Match Location:**

```
query                 MAFSQYISLAPELLLATAIFCLVFWVLRGTRTQVPKGLKSPPQ
pfam:P450                                                  <=========================================

query                 STPVVVLSGLNTIKQALVKQGDDFKGRPDLYSFTLITNGKSMTFNPDSGPVWAARRRLAQDALKSFSIASDPTSVSSCYL
pfam:P450             ==========================================================================================

query                 EEHVSKEANHLISKFQKLMAEVGHFEPVNQVVESVANVIGAMCFGKNFPRKSEEMLNLVKSSKDFVENVTSGNAVDFFPV
pfam:P450             ==========================================================================================

query                 LRYLPNPALKRFKNFNDNFVLSLQKTVQEHYQDFNKNSIQDITGALFKHSENYKDNGGLIPQEKIVNIVNDIFGAGFETV
pfam:P450             ==========================================================================================

query                 TTAIFWSILLLVTEPKVQRKIHEELDTVIGRDRQPRLSDRPQLPYLEAFILEIYRYTSFVPFTIPHSTTRDTSLNGFHIP
pfam:P450             ==========================================================================================

query                 KECCIFINQWQVNHDEKQWKDPFVFRPERFLTNDNTAIDKTLSEKVMLFGLGKRRCIGEIPAKWEVFLFLAILLHQLEFT
pat:CYTOCHROME_P450                                                   <-------->
pfam:P450             ==========================================================================================

query                 VPPGVKVDLTPSYGLTMKPRTCEHVQAWPRFSK
pfam:P450             ========================>
```

Done

Pfam: p450 – Netscape

http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF00067

Pfam: p450

Protein families database of alignments and HMMs

Wellcome Trust Sanger Institute

Home | Keyword Search | Protein Search | Browse Pfam | DNA Search | Taxonomy | ftp | Help | p450 domain

p450

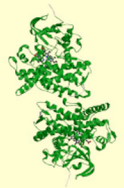


Figure 1: 2bmh Oxidoreductase(oxygenase)

Key:

Domain	Chain	Start Residue	End Residue
p450	A	5	448
p450	B	5	448

The Swissprot/PDB mapping was provided by MSD

Accession number: **PF00067**

Cytochrome P450 Add Annotation

Cytochrome P450s are involved in the oxidative degradation of various compounds. Particularly well known for their role in the degradation of environmental toxins and mutagens. Structure is mostly alpha, and binds a heme cofactor.

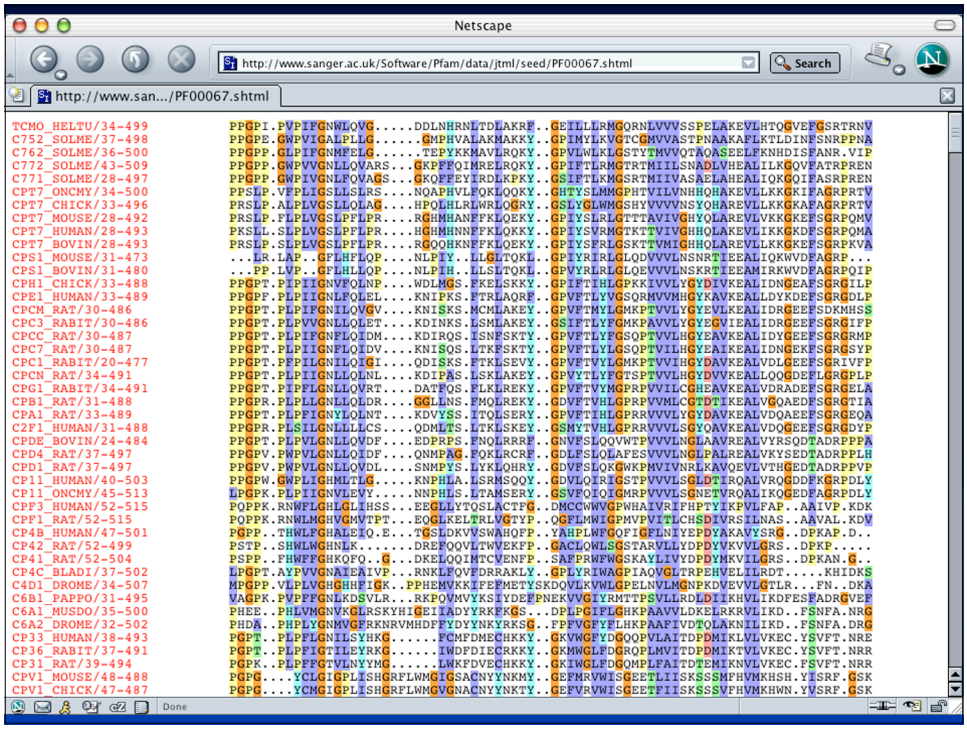
INTERPRO description (entry IPR001128)

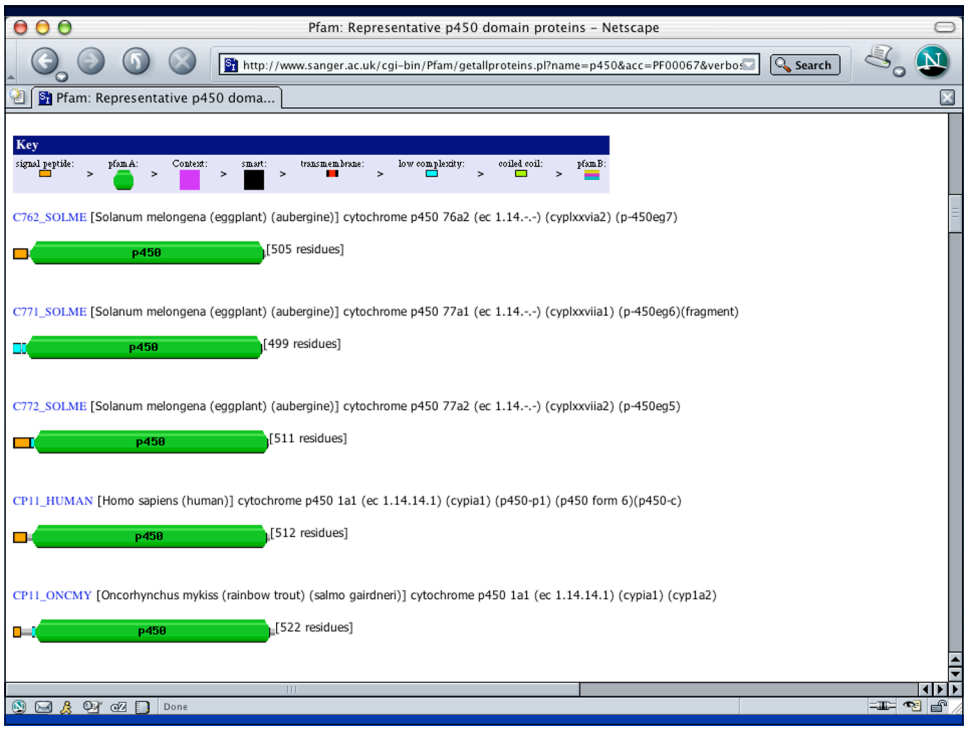
The cytochrome P450 enzymes constitute a superfamily of haem-thiolate proteins. P450 enzymes usually act as terminal oxidases in multicomponentelectron transfer chains, called P450-containing monooxygenase systems and are involved in metabolism of a plethora of both exogenous and endogenous compounds. P450-containing monooxygenase systems primarily fall into two major classes: bacterial/mitochondrial (type I), and microsomal (type II). All P450 enzymes can be categorised into two main groups, the so-called B- and E-classes: P450 proteins of prokaryotic 3-component systems and fungal P450nor (CYP55) belong to the B-class; all other known P450 proteins from distinct systems are of the E-class [MEDLINE:93135827].

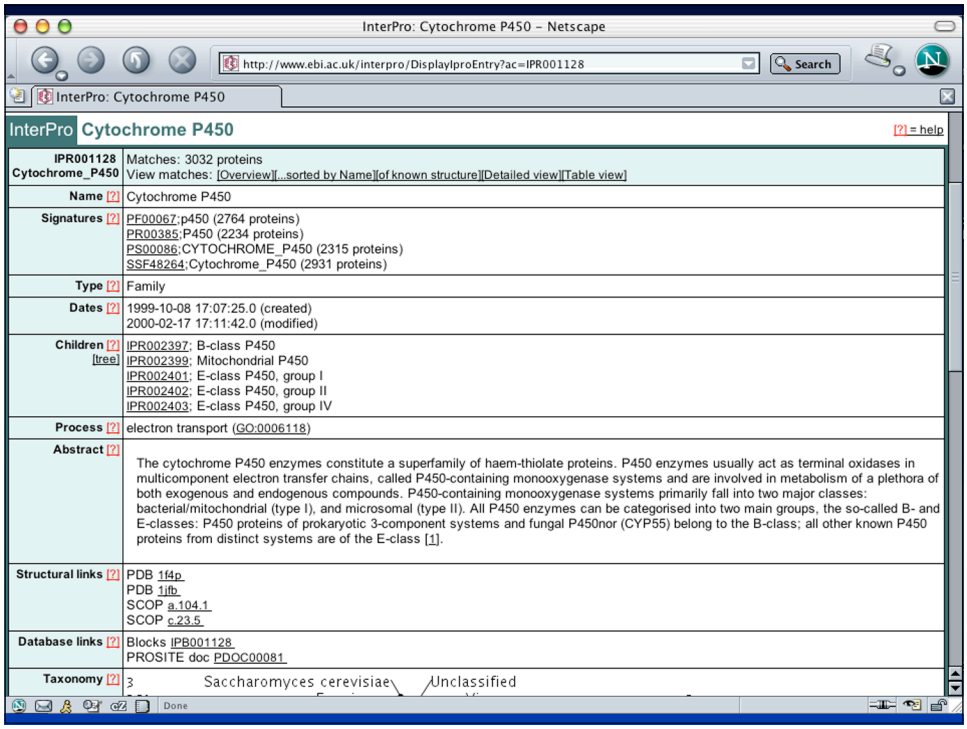
QuickGO	
PROCESS :	electron transport (GO:0006118)

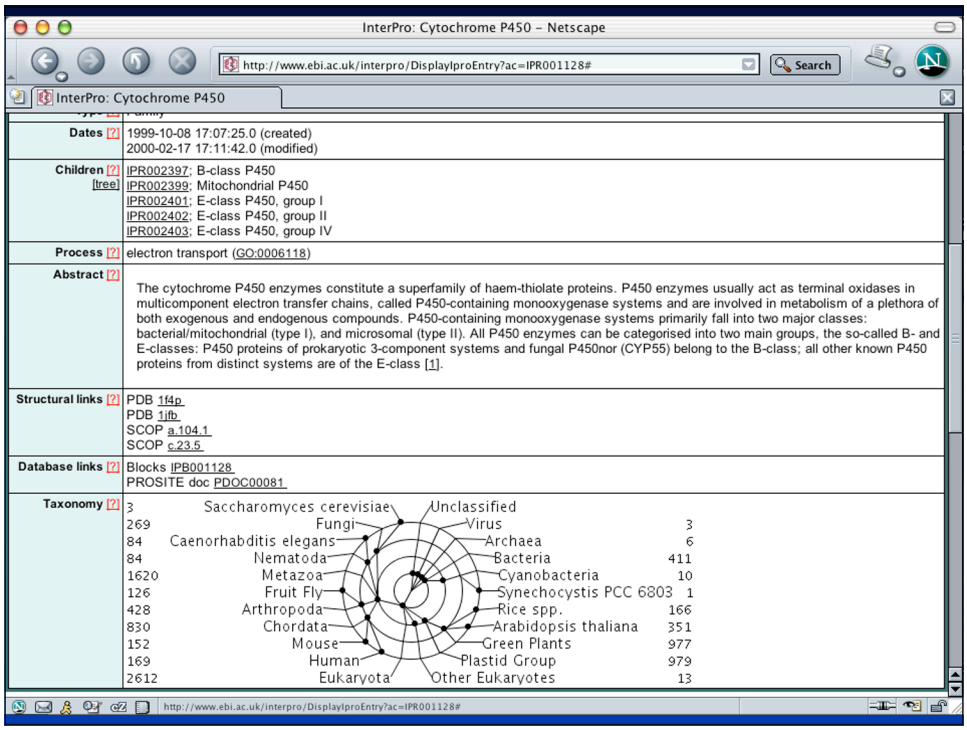
→ Alignments
→ Domain Structure

For additional annotation, see the PROSITE document PDOC00081 [Expasy | SRS-UK | SRS-USA]


Netscape
http://www.sanger.ac.uk/Software/Pfam/data/jtml/seed/PF00067.shtml
Search
http://www.san.../PF00067.shtml
Done

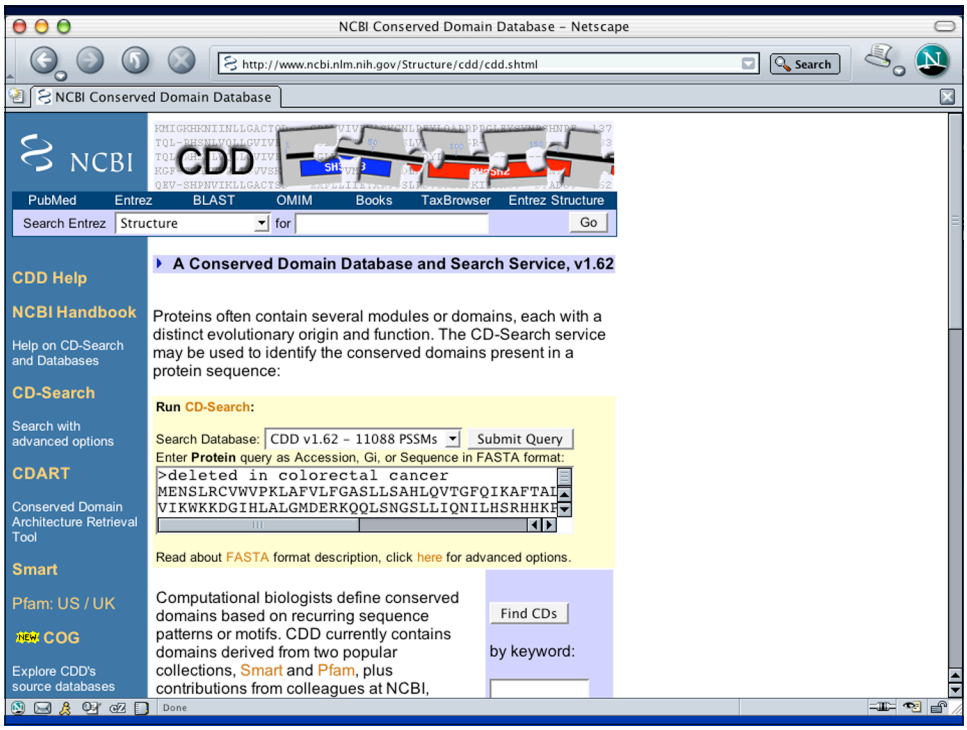

Pfam: Representative p450 domain proteins - Netscape
http://www.sanger.ac.uk/cgi-bin/Pfam/getallproteins.pl?name=p450&acc=PF00067&verbos
Search
Pfam: Representative p450 doma...
Key
signal peptide:
pfamA:
Context:
smart:
transmembrane:
low complexity:
coiled coil:
pfamB:
C762_SOLME [Solanum melongena (eggplant) (aubergine)] cytochrome p450 76a2 (ec 1.14.-.-) (cyplxxvia2) (p-450eg7)
p450
[505 residues]
C771_SOLME [Solanum melongena (eggplant) (aubergine)] cytochrome p450 77a1 (ec 1.14.-.-) (cyplxxviia1) (p-450eg6)(fragment)
p450
[499 residues]
C772_SOLME [Solanum melongena (eggplant) (aubergine)] cytochrome p450 77a2 (ec 1.14.-.-) (cyplxxviia2) (p-450eg5)
p450
[511 residues]
CP11_HUMAN [Homo sapiens (human)] cytochrome p450 1a1 (ec 1.14.14.1) (cypia1) (p450-p1) (p450 form 6)(p450-c)
p450
[512 residues]
CP11_ONCMY [Oncorhynchus mykiss (rainbow trout) (salmo gairdneri)] cytochrome p450 1a1 (ec 1.14.14.1) (cypia1) (cyp1a2)
p450
[522 residues]
Done


InterPro: Cytochrome P450 - Netscape
http://www.ebi.ac.uk/interpro/DisplayIproEntry?ac=IPR001128
Search
InterPro: Cytochrome P450
InterPro Cytochrome P450
[?] = help
IPR001128 Cytochrome_P450
Matches: 3032 proteins
View matches: [Overview][...sorted by Name][of known structure][Detailed view][Table view]
Name [?]
Cytochrome P450
Signatures [?]
PF00067;p450 (2764 proteins)
PR00385;P450 (2234 proteins)
PS00086;CYTOCHROME_P450 (2315 proteins)
SSF48264;Cytochrome_P450 (2931 proteins)
Type [?]
Family
Dates [?]
1999-10-08 17:07:25.0 (created)
2000-02-17 17:11:42.0 (modified)
Children [?] [tree]
IPR002397; B-class P450
IPR002399; Mitochondrial P450
IPR002401; E-class P450, group I
IPR002402; E-class P450, group II
IPR002403; E-class P450, group IV
Process [?]
electron transport (GO:0006118)
Abstract [?]
The cytochrome P450 enzymes constitute a superfamily of haem-thiolate proteins. P450 enzymes usually act as terminal oxidases in multicomponent electron transfer chains, called P450-containing monooxygenase systems and are involved in metabolism of a plethora of both exogenous and endogenous compounds. P450-containing monooxygenase systems primarily fall into two major classes: bacterial/mitochondrial (type I), and microsomal (type II). All P450 enzymes can be categorised into two main groups, the so-called B- and E-classes: P450 proteins of prokaryotic 3-component systems and fungal P450nor (CYP55) belong to the B-class; all other known P450 proteins from distinct systems are of the E-class [1].
Structural links [?]
PDB 1f4p
PDB 1jfb
SCOP a.104.1
SCOP c.23.5
Database links [?]
Blocks IPB001128
PROSITE doc PDOC00081
Taxonomy [?]
3 Saccharomyces cerevisiae Unclassified
Done


InterPro: Cytochrome P450 - Netscape
http://www.ebi.ac.uk/interpro/DisplayIproEntry?ac=IPR001128#
Search
InterPro: Cytochrome P450
Dates [?]
1999-10-08 17:07:25.0 (created)
2000-02-17 17:11:42.0 (modified)
Children [?] [tree]
IPR002397; B-class P450
IPR002399; Mitochondrial P450
IPR002401; E-class P450, group I
IPR002402; E-class P450, group II
IPR002403; E-class P450, group IV
Process [?]
electron transport (GO:0006118)
Abstract [?]
The cytochrome P450 enzymes constitute a superfamily of haem-thiolate proteins. P450 enzymes usually act as terminal oxidases in multicomponent electron transfer chains, called P450-containing monooxygenase systems and are involved in metabolism of a plethora of both exogenous and endogenous compounds. P450-containing monooxygenase systems primarily fall into two major classes: bacterial/mitochondrial (type I), and microsomal (type II). All P450 enzymes can be categorised into two main groups, the so-called B- and E-classes: P450 proteins of prokaryotic 3-component systems and fungal P450nor (CYP55) belong to the B-class; all other known P450 proteins from distinct systems are of the E-class [1].
Structural links [?]
PDB 1f4p
PDB 1jfb
SCOP a.104.1
SCOP c.23.5
Database links [?]
Blocks IPB001128
PROSITE doc PDOC00081
Taxonomy [?]
3 Saccharomyces cerevisiae Unclassified
269 Fungi
84 Caenorhabditis elegans
84 Nematoda
1620 Metazoa
126 Fruit Fly
428 Arthropoda
830 Chordata
152 Mouse
169 Human
2612 Eukaryota
Virus 3
Archaea 6
Bacteria 411
Cyanobacteria 10
Synechocystis PCC 6803 1
Rice spp. 166
Arabidopsis thaliana 351
Green Plants 977
Plastid Group 979
Other Eukaryotes 13
http://www.ebi.ac.uk/interpro/DisplayIproEntry?ac=IPR001128#

Conserved Domain Database (CDD)

- Identify conserved domains in a protein sequence
- "Secondary database"
  - Pfam A and B
  - Simple Modular Architecture Research Tool (SMART)
- Search performed using RPS-BLAST
  - Query sequence is used to search a database of precalculated position-specific scoring tables
  - *Not* the same method used by ProfileScan
- *http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml*



NCBI CD-Search - Netscape

http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

NCBI CD-Search

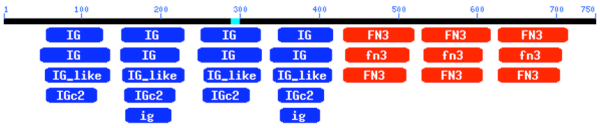
NCBI

NCBI Conserved Domain Search

New Search | PubMed | Nucleotide | Protein | Structure | CDD | Taxonomy | Help?

```
RPS-BLAST 2.2.6 [Apr-09-2003]

Query= local sequence: DELETED IN COLORECTAL CANCER
       (750 letters)

Database: #cdd.v1.62
          11,088 PSSMs; 2,717,223 total columns
```

Masked-out region, low complexity



Show | Domain Relatives

.. This CD alignment includes 3D structure. To display structure, download **Cn3D**!

PSSMs producing significant alignments:	Score (bits)	E value
gnl|CDD|14799 cd00063, FN3, Fibronectin type 3 domain; One of three types of...	73.4	9e-14
gnl|CDD|14799 cd00063, FN3, Fibronectin type 3 domain; One of three types of...	72.7	2e-13
gnl|CDD|14799 cd00063, FN3, Fibronectin type 3 domain; One of three types of...	68.0	3e-12

http://www.ncbi.nlm.nih.gov/BLAST/filtered.html

NCBI CD-Search - Netscape

http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi#ali2078917053

NCBI CD-Search

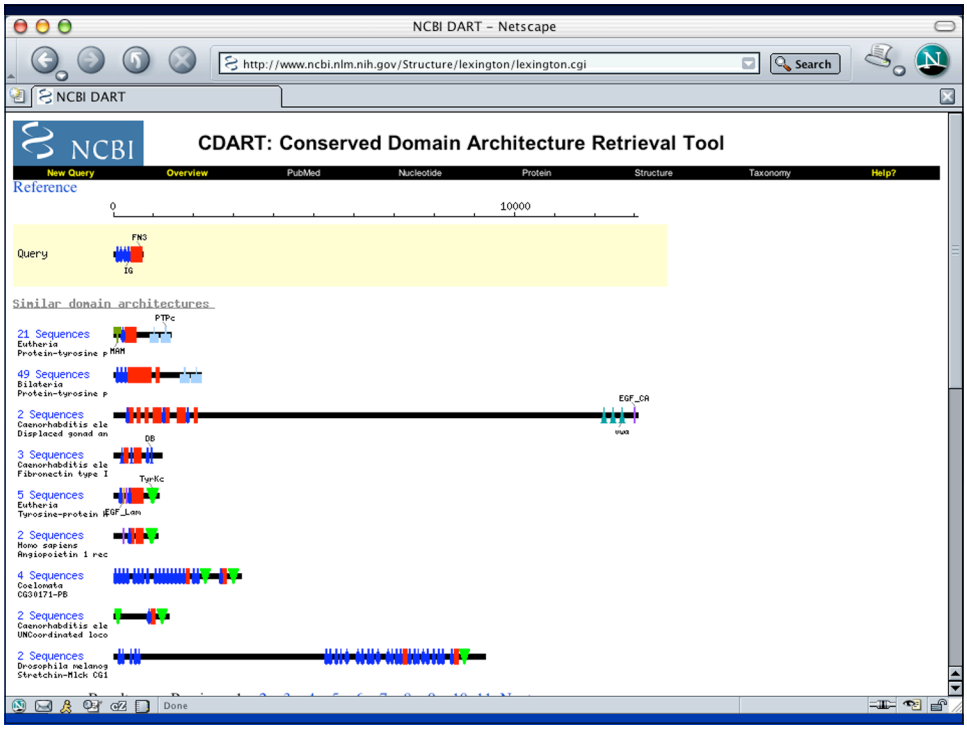
gnl|CDD|14799, cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found in the plasma protein fibronectin. Its tenth fibronectin type III repeat contains an RGD cell recognition sequence in a flexible loop between 2 strands. Approximately 2% of all animal proteins contain the FN3 repeat; including extracellular and intracellular proteins, membrane spanning cytokine receptors, growth hormone receptors, tyrosine phosphatase receptors, and adhesion molecules. FN3-like domains are also found in bacterial glycosyl hydrolases.

```
             CD-Length = 93 residues, 100.0% aligned
             Score = 73.4 bits (179), Expect = 9e-14

Query:  429  PSAPRDVVPVLVSSRFVRLSWRPPAEAKGNIQTFTVFFSREGDNRERALNTTQPGSLQLT  488
Sbjct:  1    PSPPTNLRVTDVTSTSVTLSWTPPEDDGGPITGYVVEYREKGSGDWKEVEVTPGSETSYT  60

Query:  489  VGNLKPEAMYTFRVVAYNEWGPGESSQPIKVAT  521
Sbjct:  61   LTGLKPGTEYEFRVRAVNGGGESPPSESVTVTT  93
```

gnl|CDD|14799, cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found in the plasma protein fibronectin. Its tenth fibronectin type III repeat contains an RGD cell recognition sequence in a flexible loop between 2 strands. Approximately 2% of all animal proteins contain the FN3 repeat; including extracellular and intracellular proteins, membrane spanning cytokine receptors, growth hormone receptors, tyrosine phosphatase receptors, and adhesion molecules. FN3-like domains are also found in bacterial glycosyl hydrolases.

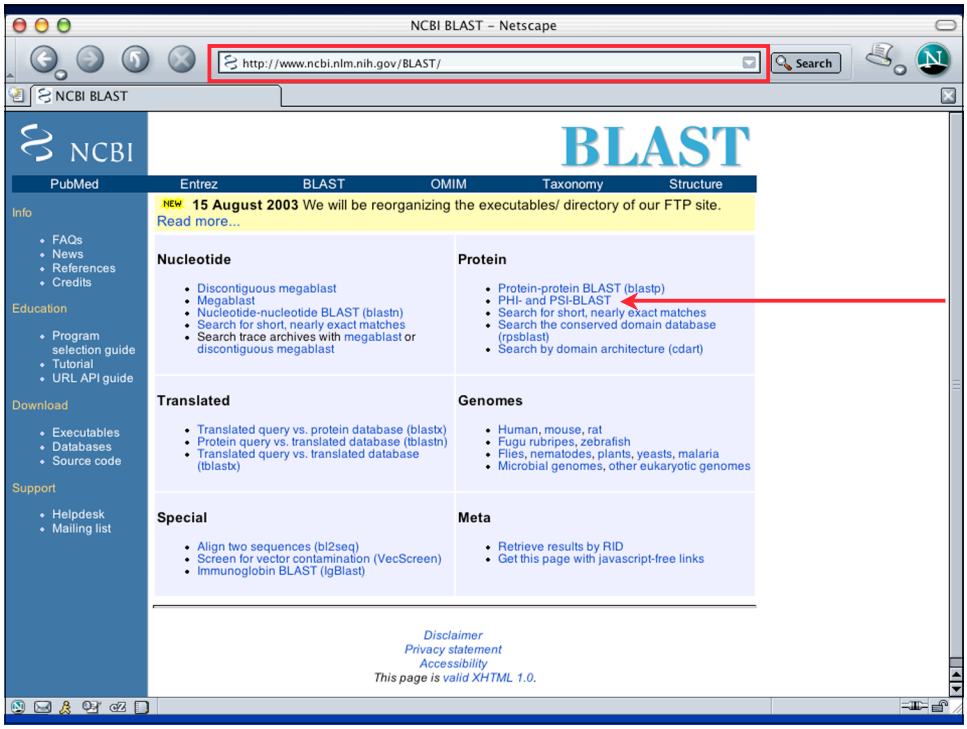
```
             CD-Length = 93 residues, 100.0% aligned
             Score = 72.7 bits (177), Expect = 2e-13

Query:  528  PGPVENLQAVSTSPTSILITWEPPAYANGPVQGYRLFCTEVSTGKEQNIEV---DGLSYK  584
Sbjct:  1    PSPPTNLRVTDVTSTSVTLSWTPPEDDGGPITGYVVEYREKGSGDWKEVEVTPGSETSYT  60

Query:  585  LEGLKKFTEYSLRFLAYNRYGPGVSTDDITVVT  617
Sbjct:  61   LTGLKPGTEYEFRVRAVNGGGESPPSESVTVTT  93
```

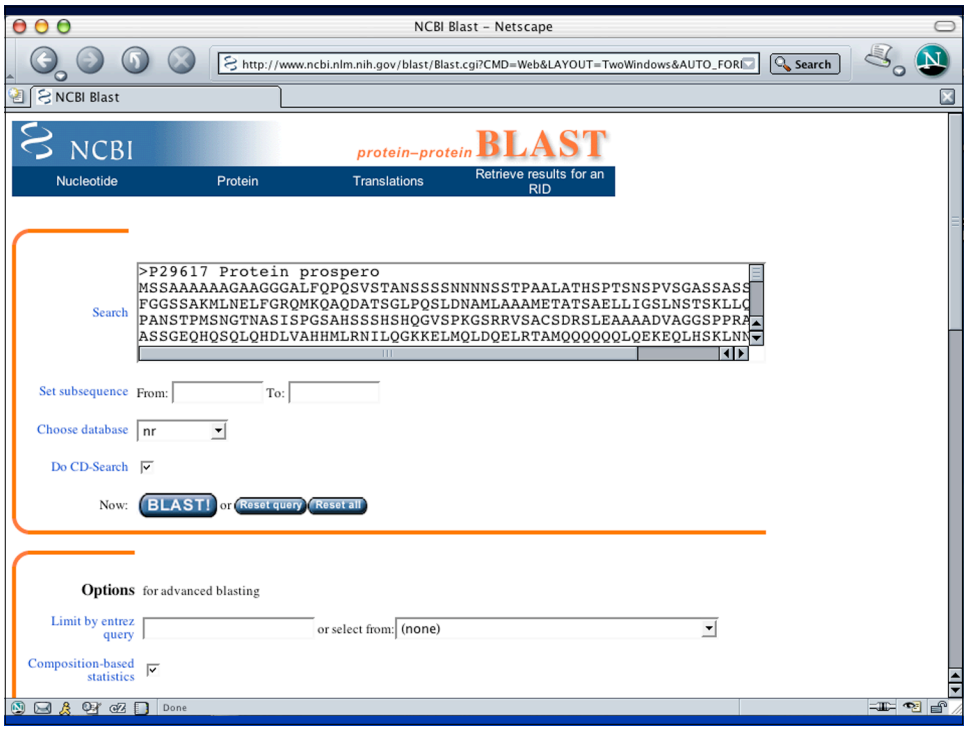
gnl|CDD|14799, cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found in the plasma protein fibronectin. Its

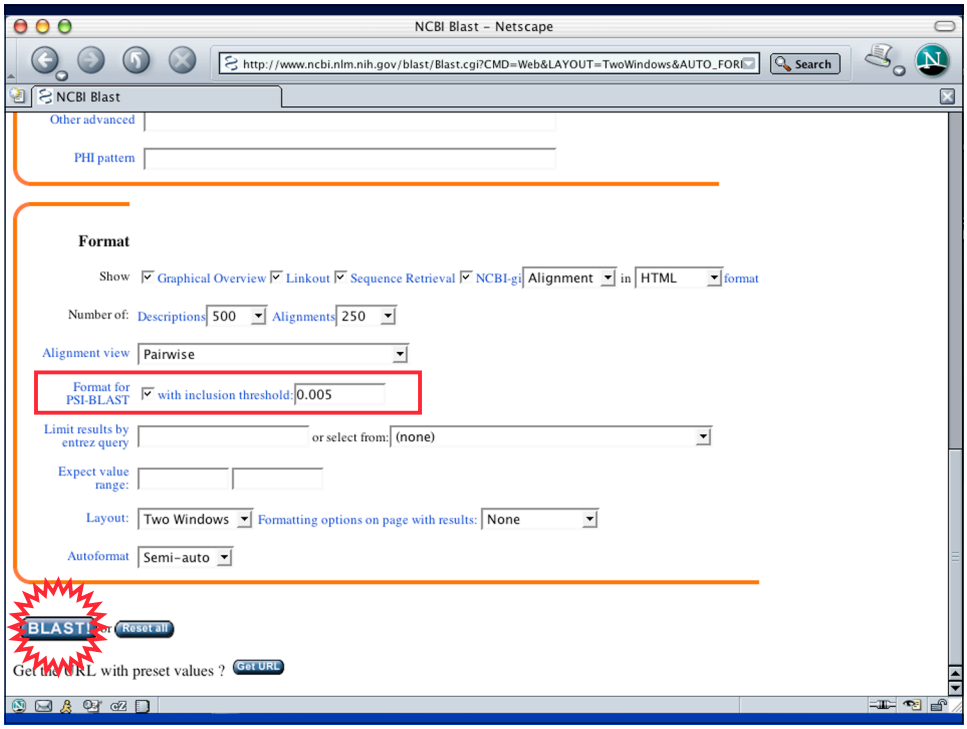


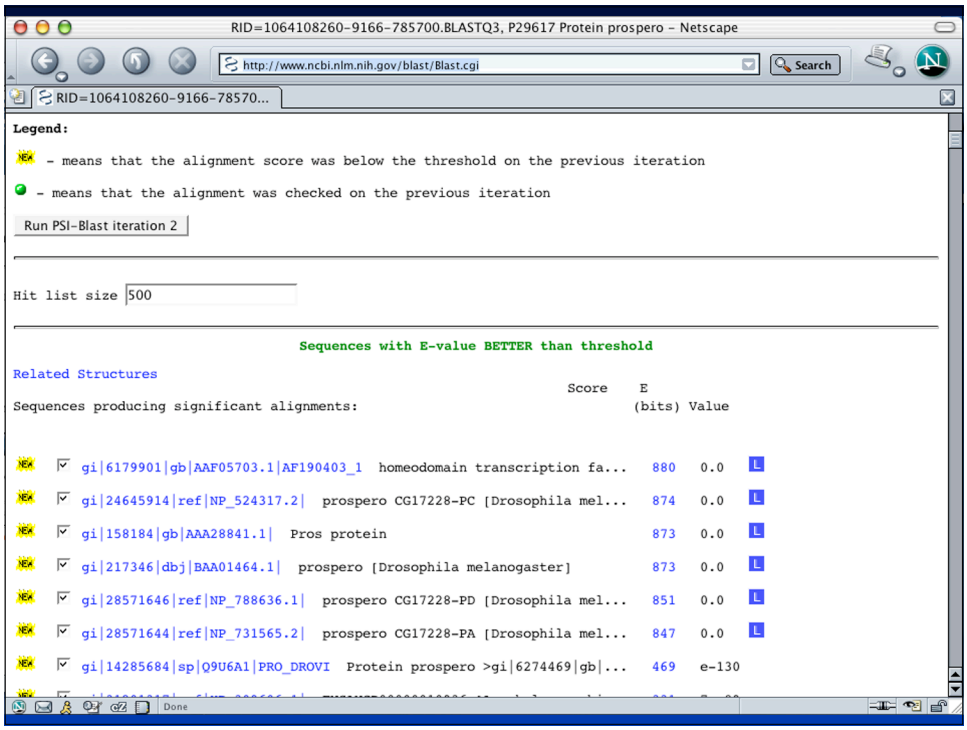
PSI-BLAST

- Position-Specific Iterated BLAST search
- Easy-to-use version of a profile-based search
  - Perform BLAST search against protein database
  - Use results to calculate a position-specific scoring matrix
  - PSSM replaces query for next round of searches
  - May be iterated until no new significant alignments are found
    - Convergence – all related sequences deemed found
    - Divergence – query is too broad, make cutoffs more stringent


NCBI BLAST - Netscape
http://www.ncbi.nlm.nih.gov/BLAST/
Search
NCBI BLAST
NCBI
BLAST
PubMed
Entrez
BLAST
OMIM
Taxonomy
Structure
Info
FAQs
News
References
Credits
Education
Program selection guide
Tutorial
URL API guide
Download
Executables
Databases
Source code
Support
Helpdesk
Mailing list
NEW 15 August 2003 We will be reorganizing the executables/ directory of our FTP site.
Read more...
Nucleotide
Discontiguous megablast
Megablast
Nucleotide-nucleotide BLAST (blastn)
Search for short, nearly exact matches
Search trace archives with megablast or discontiguous megablast
Protein
Protein-protein BLAST (blastp)
PHI- and PSI-BLAST
Search for short, nearly exact matches
Search the conserved domain database (rpsblast)
Search by domain architecture (cdart)
Translated
Translated query vs. protein database (blastx)
Protein query vs. translated database (tblastn)
Translated query vs. translated database (tblastx)
Genomes
Human, mouse, rat
Fugu rubripes, zebrafish
Flies, nematodes, plants, yeasts, malaria
Microbial genomes, other eukaryotic genomes
Special
Align two sequences (bl2seq)
Screen for vector contamination (VecScreen)
Immunoglobin BLAST (IgBlast)
Meta
Retrieve results by RID
Get this page with javascript-free links
Disclaimer
Privacy statement
Accessibility
This page is valid XHTML 1.0.


NCBI Blast - Netscape
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORM
Search
NCBI Blast
NCBI
protein-protein BLAST
Nucleotide
Protein
Translations
Retrieve results for an RID
Search
>P29617 Protein prospero
MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSASS
FGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQ
PANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSACSDRSLEAAAADVAGGSPPRA
ASSGEQHQSQLQHDLVAHHMLRNILQGKKELMQLDQELRTAMQQQQQQLQEKEQLHSKLNN
Set subsequence
From:
To:
Choose database
nr
Do CD-Search
Now:
BLAST!
or
Reset query
Reset all
Options
for advanced blasting
Limit by entrez query
or select from:
(none)
Composition-based statistics
Done


NCBI Blast - Netscape
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORM
Search
NCBI Blast
Other advanced
PHI pattern
Format
Show
Graphical Overview
Linkout
Sequence Retrieval
NCBI-gi
Alignment
in
HTML
format
Number of:
Descriptions
500
Alignments
250
Alignment view
Pairwise
Format for PSI-BLAST
with inclusion threshold:
0.005
Limit results by entrez query
or select from:
(none)
Expect value range:
Layout:
Two Windows
Formatting options on page with results:
None
Autoformat
Semi-auto
BLAST!
Reset all
Get the URL with preset values ?
Get URL


RID=1064108260-9166-785700.BLASTQ3, P29617 Protein prospero - Netscape
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Search
RID=1064108260-9166-78570...
Legend:
- means that the alignment score was below the threshold on the previous iteration
- means that the alignment was checked on the previous iteration
Run PSI-Blast iteration 2
Hit list size
500
Sequences with E-value BETTER than threshold
Related Structures
Score
E
Sequences producing significant alignments:
(bits)
Value
gi|6179901|gb|AAF05703.1|AF190403_1 homeodomain transcription fa... 880 0.0
gi|24645914|ref|NP_524317.2| prospero CG17228-PC [Drosophila mel... 874 0.0
gi|158184|gb|AAA28841.1| Pros protein 873 0.0
gi|217346|dbj|BAA01464.1| prospero [Drosophila melanogaster] 873 0.0
gi|28571646|ref|NP_788636.1| prospero CG17228-PD [Drosophila mel... 851 0.0
gi|28571644|ref|NP_731565.2| prospero CG17228-PA [Drosophila mel... 847 0.0
gi|14285684|sp|Q9U6A1|PRO_DROVI Protein prospero >gi|6274469|gb|... 469 e-130
Done

RID=1064108260-9166-785700.BLASTQ3, P29617 Protein prospero - Netscape

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

RID=1064108260-9166-78570...

gi|28521642|ref|XP_127011.2|	RIKEN cDNA 1700058C01 [Mus musculus...	193	2e-47
gi|3372869|gb|AAC28353.1|	Prox1 [Xenopus laevis]	190	2e-46
gi|27680210|ref|XP_223067.1|	similar to prospero-related homeobo...	187	2e-45
gi|1117962|gb|AAC59781.1|	prospero_like protein	159	5e-37
gi|21753053|dbj|BAC04278.1|	unnamed protein product [Homo sapiens]	146	3e-33
gi|11071926|dbj|BAB17311.1|	Prox 1 [Cynops pyrrhogaster]	143	2e-32
gi|4809335|gb|AAD30180.1|AC006530_2	homeobox prospero-like prote...	87	2e-15
gi|7512234|pir||JC5496	Prox 1 protein 671 - chicken	70	3e-10

Run PSI-Blast iteration 2

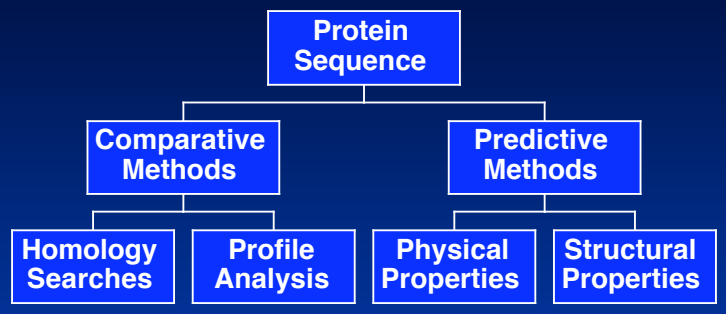
… 3 … 4 … *n* … "Converged"

Sequences with E-value WORSE than threshold

gi|6466795|gb|AAF13029.1|AF070733_1	transcription factor Prox1 [...	41	0.21
gi|15923965|ref|NP_371499.1|	ClpB chaperone homologue [Staphyloc...	35	8.6
gi|21282586|ref|NP_645674.1|	ClpB chaperone homologue-ORFID:MW08...	35	8.7
gi|27467592|ref|NP_764229.1|	clpB protein [Staphylococcus epider...	35	8.8

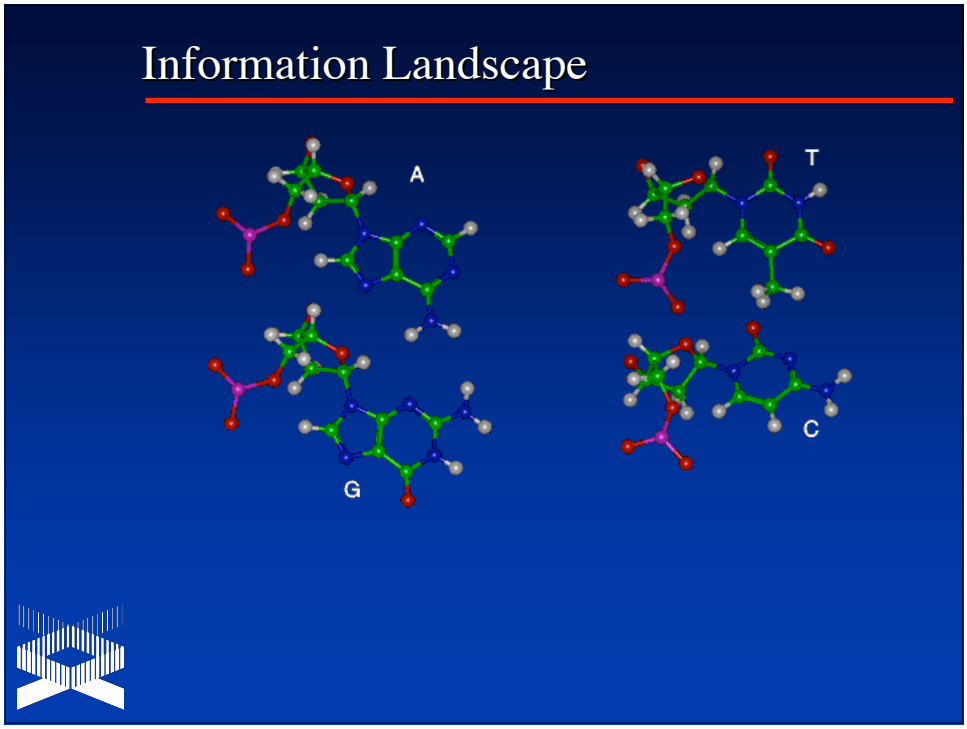
Run PSI-Blast iteration 2

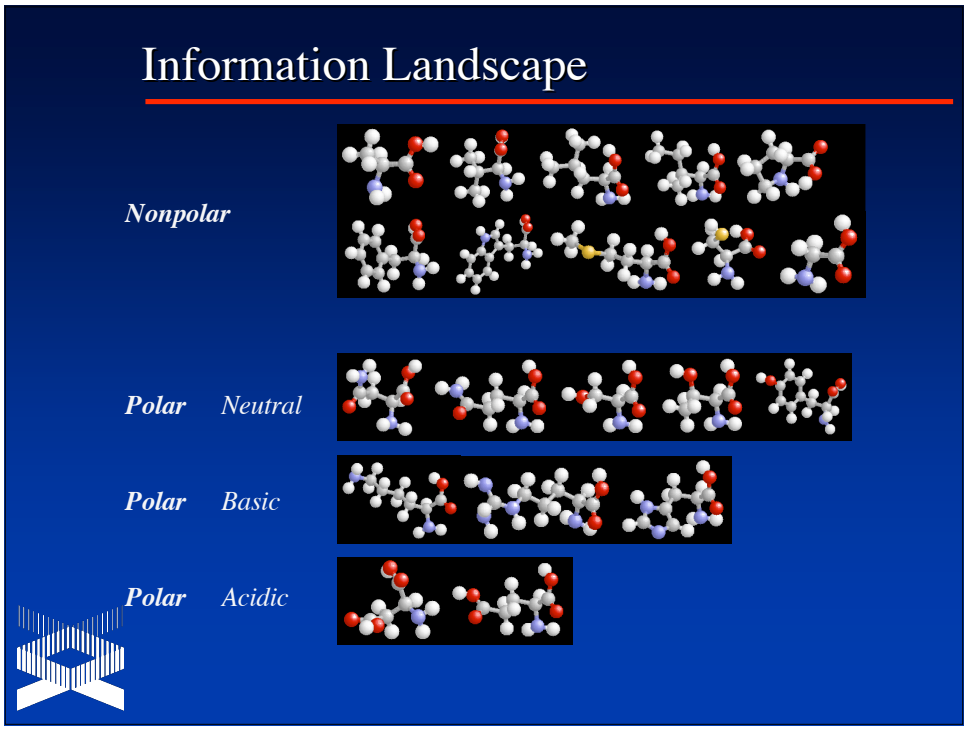
Protein Sequence Analysis



- *Common structure?*
- *Common function?*
- *Evolutionary relationship?*
- *Global or local similarity?*

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


Information Landscape
A
T
G
C


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://www.expasy.ch/tools/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

Rank	ID	DIST	LEN2	POS1	POS2	pI	DE
1	>p1;s18193	0.00	727	1	727	5.33	autoantigen NOR-90 - human
2	ubf1_human	1.36	764	1	764	5.62	NUCLEOLAR TRANSCRIPTION FACTOR 1
3	ubf1_mouse	1.40	765	1	765	5.55	NUCLEOLAR TRANSCRIPTION FACTOR 1
4	ubf1_rat	1.57	764	1	764	5.61	NUCLEOLAR TRANSCRIPTION FACTOR 1
5	ubf1_xenla	3.95	677	1	677	5.79	NUCLEOLAR TRANSCRIPTION FACTOR 1
6	ubf2_xenla	4.18	701	1	701	6.05	NUCLEOLAR TRANSCRIPTION FACTOR 2
7	>p1;s57552	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
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

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	ubf2_mla	4.18	701	1	701	6.05	NUCLEOLAR TRANSCRIPTION FACTOR 2
7	>p1;s57152	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
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

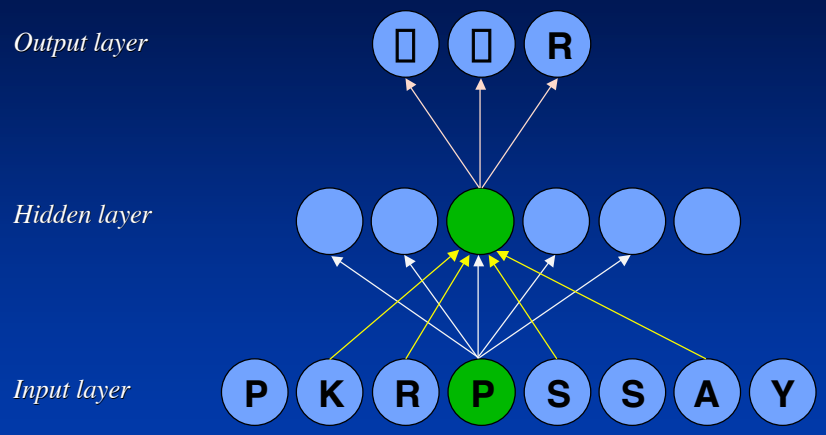
Secondary Structure Prediction

- Deduce the most likely position of alpha-helices and beta-strands
- Confirm structural or functional relationships when sequence similarity is weak
- Determine guidelines for rational selection of specific mutants for further laboratory study
- Basis for further structure-based studies

Folding Classes

		+	/
Cyt c	*CD4*	*Staph nuclease*	*Triose phosphate isomerase*
Globins Orthogonal EF-hand Up-Down Cytochrome	Orthogonal Super-barrel Greek key Sandwich Jelly roll	Split sandwich Meander Metal-rich Open roll OB/UB roll	TIM barrel Doubly-wound

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

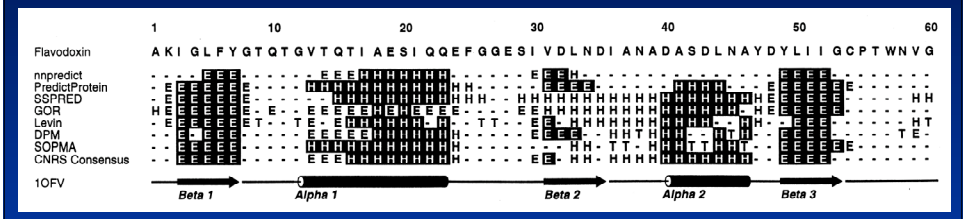
```
Joe Buzzcut
National Human Genome Research Institute, NIH
buzzcut@nhgri.nih.gov
# flavodoxin - Anacystis nidulans
AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADASDLNAYDYLIIGCPTWNVGELQSDWEGIY
DDLDSVNFQGKKVAYFGAGDQVGYSDNFQDAMGILEEKISSLGSQTVGYWPIEGYDFNESKAVRNNQFVG
LAIDEDNQPDLTKNRIKTWVSQLKSEFGL
```

Secondary structure

```
          ....,....1....,....2....,....3....,....4....,....5....,....6
AA       |AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADASDLNAYDYLIIGCPTWNVG|
PHD sec  | EEEEEEE    HHHHHHHHHHHH     EEEEE HHH HHHH    EEEEE        |
Rel sec  |938999736982489999999999767982443213241278631241999861547765|
Detail:
prH sec  |000000000014689999999999821000011112565388764321000001111111|
prE sec  |058998852000000000000000000000366554210000000001489987412000 2|
prL sec  |931000137985310000000000178985222344324511234554000114667776|
```

- *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

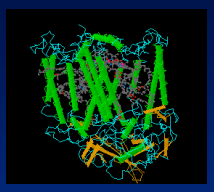
```
>sp|P05019|IGFB_HUMAN INSULIN-LIKE GROWTH FACTOR IB PRECURSOR
MGKISSLPTQLFKCCFCDFLKVKMHTMSSSHLFYLALCLLTFTSSATAGPETLCGAELVDALQFVCGDRG
```

N-terminal end only
Eukaryotic set

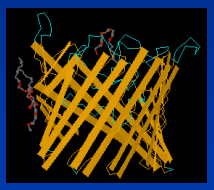
```
************************* SignalP predictions *************************
Using networks trained on euk data

>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

Transmembrane Classes



- Helix bundles
  Long stretches of apolar amino acids
  Fold into transmembrane alpha-helices
  "Positive-inside rule"
  - *Cell surface receptors*
  - *Ion channels*
  - *Active and passive transporters*



- Beta-barrel
  Anti-parallel sheets rolled into cylinder
  - *Outer membrane of Gram-negative bacteria*
  - *Porins (passive, selective diffusion)*

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...
```

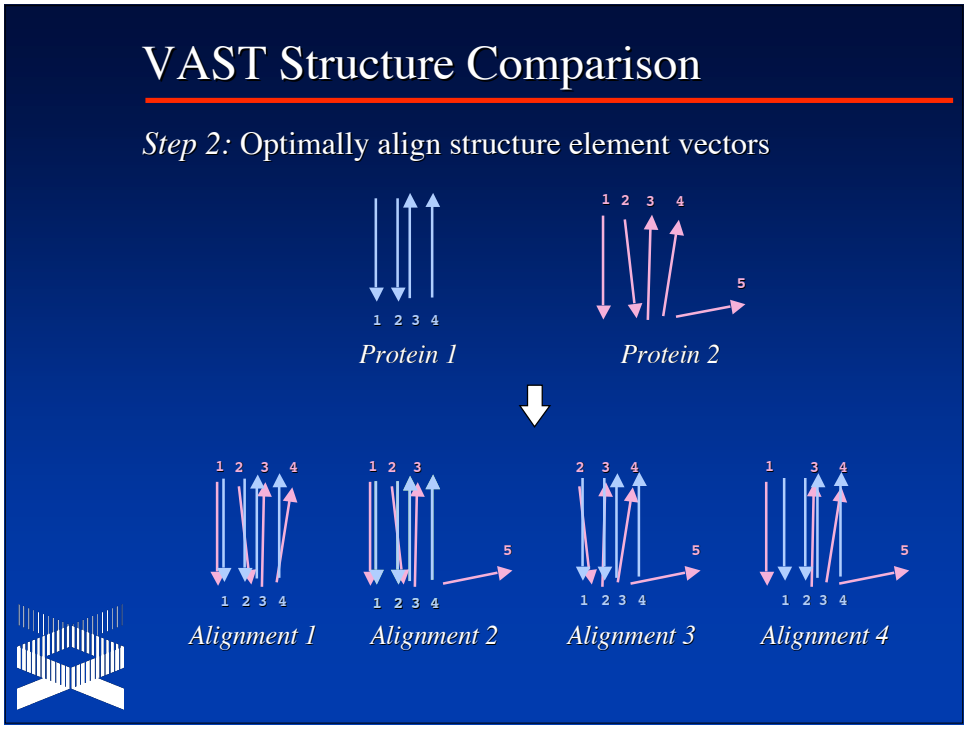
```
                 ....,....37...,....38...,....39...,....40...,....41...,....42
        AA      |YSLKYDYPLDGNQELIALGLGNIVCGVFRGFAGSTALSRSAVQESTGGKTQIAGLIGAII|
        PHD htm |                      HHHHHHHHHHHHH              HHHHHHHHH|
        Rel htm |3688999999999999986411046677776554312577788877776214677888 8|
 detail:        |                                                            |
        prH htm |310000000000000000124457888888877765321110000111135788899999|
        prL htm |689999999999999999875542111111122234678889999888864211100000|
                                         .
                                         .
                                         .
        PHDThtm |iiiiiiiiiiiiiiiiiiiTTTTTTTTTTTTTTTTTTooooooooooooooooTTTTTTTT|
```

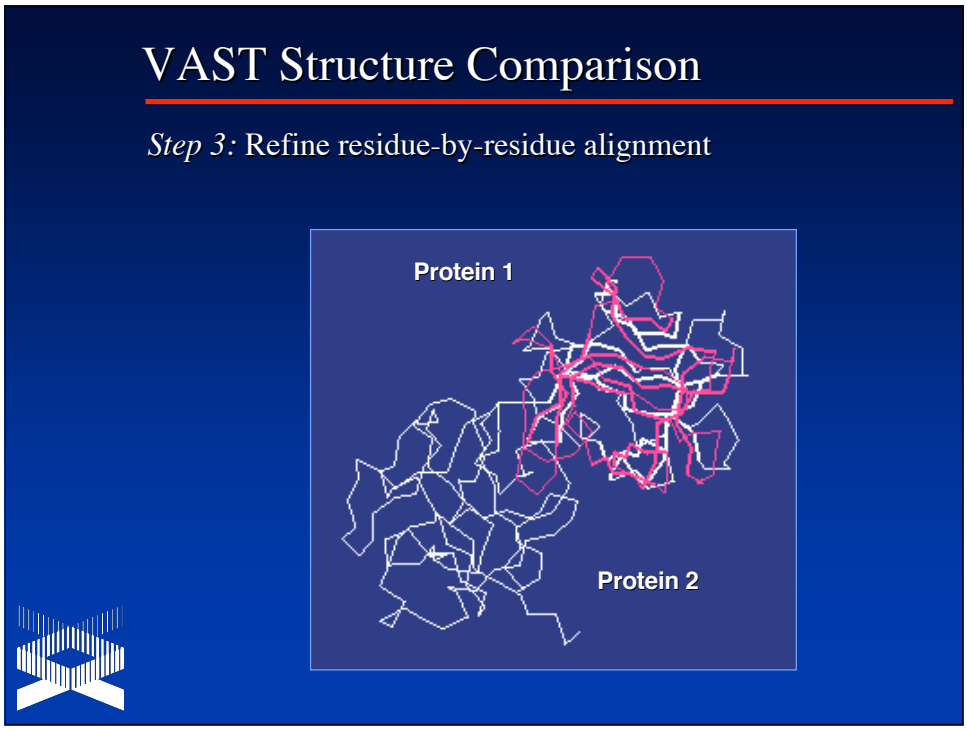
Predicting Tertiary Structure

- Sequence specifies conformation, *but* conformation does *not* specify sequence
- Structure is conserved to a much greater extent than sequence
  - Limited number of protein folds
- Similarities between proteins may not necessarily be detected through "traditional" methods

VAST Structure Comparison

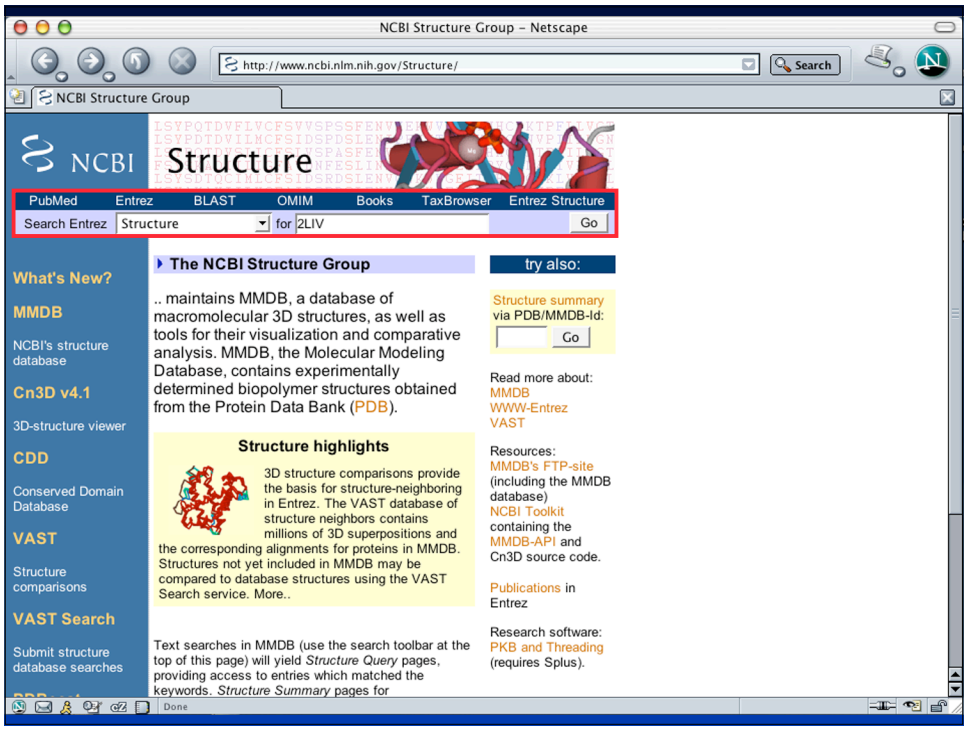
Step 1: Construct vectors for secondary structure elements

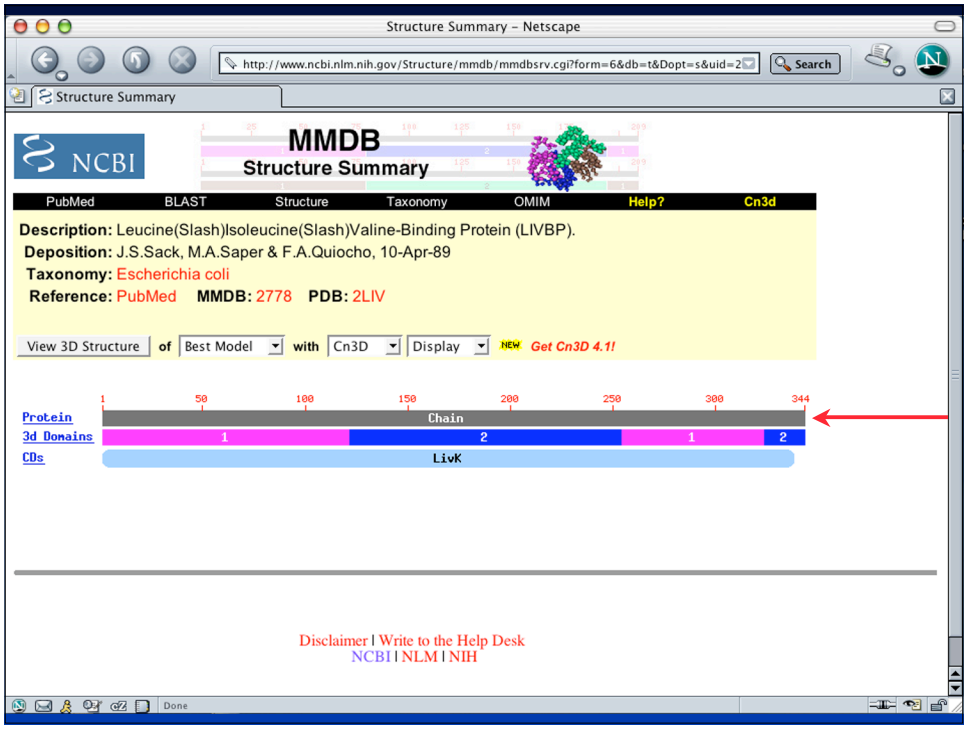

VAST Structure Comparison
Step 2: Optimally align structure element vectors
1 2 3 4
Protein 1
1 2 3 4
5
Protein 2
1 2 3 4
1 2 3 4
Alignment 1
1 2 3
5
1 2 3 4
Alignment 2
2 3 4
5
1 2 3 4
Alignment 3
1 3 4
5
1 2 3 4
Alignment 4

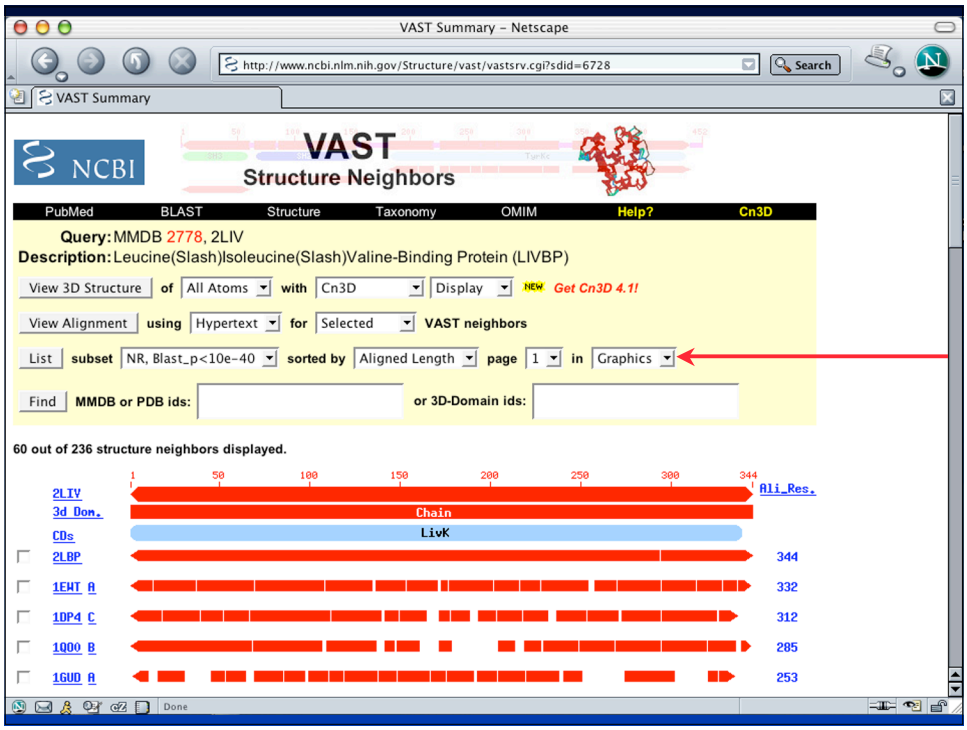

VAST Structure Comparison
Step 3: Refine residue-by-residue alignment
Protein 1
Protein 2

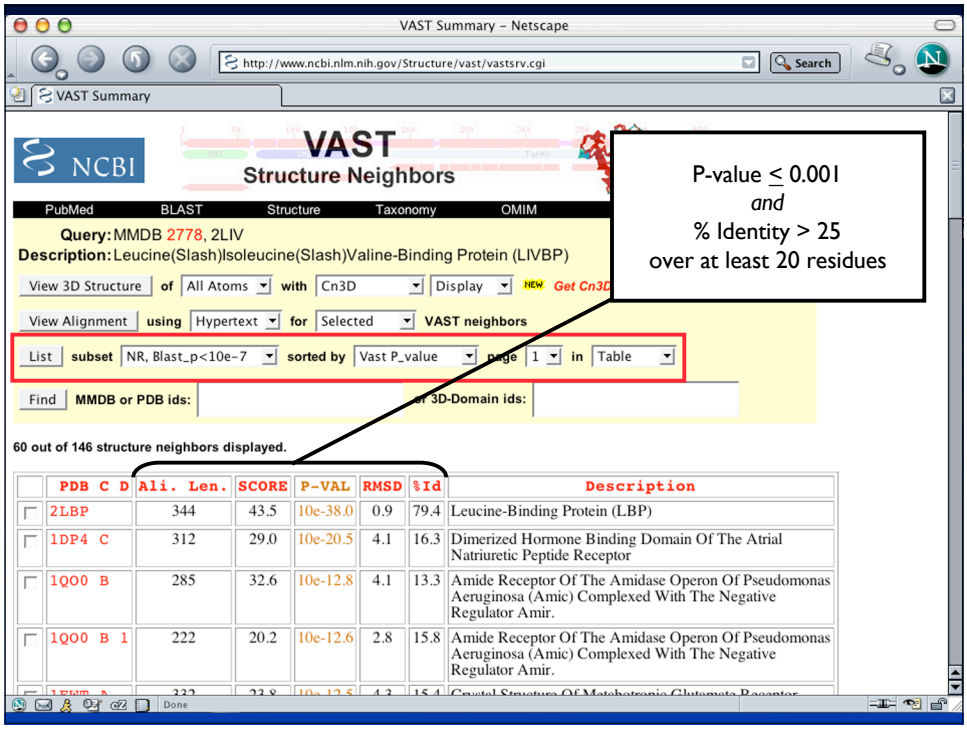
VAST Shortcomings

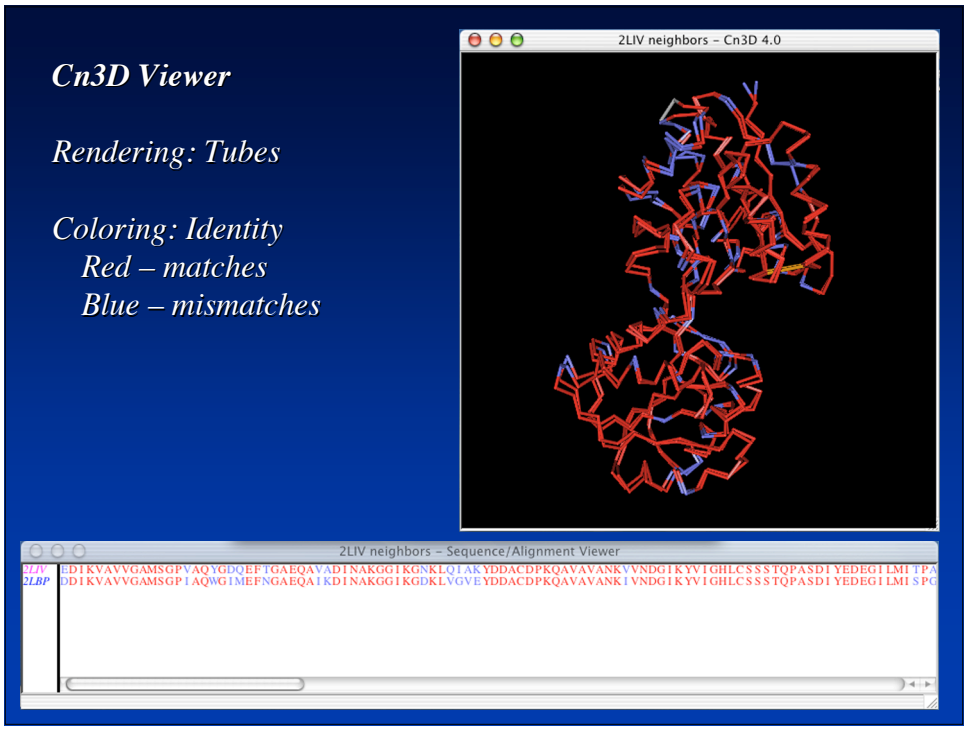
- Not the best method for determining structural similarities
- Reducing a structure to a series of vectors necessarily results in a loss of information (less confidence in prediction)
- Regardless of the "simplicity" of the method, provides a simple and fast first answer to the question of structural similarity

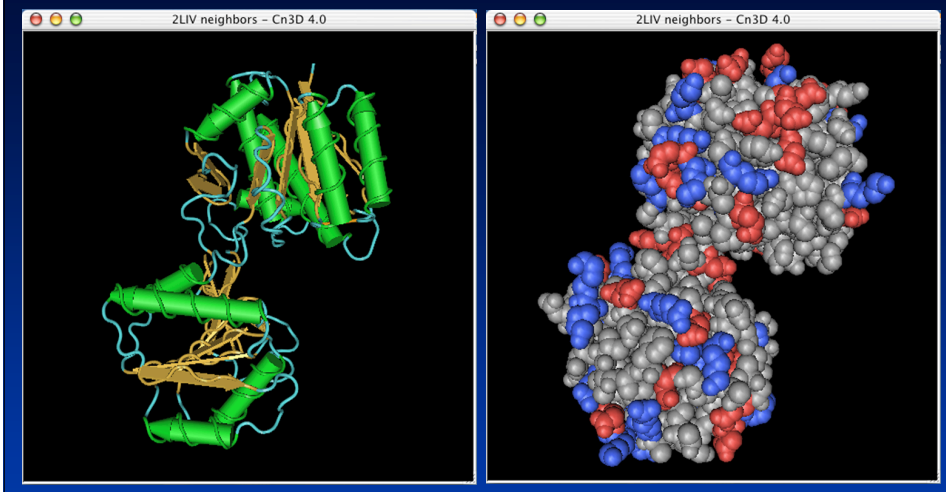



Structure Summary – Netscape
http://www.ncbi.nlm.nih.gov/Structure/mmdb/mmdbsrv.cgi?form=6&db=t&Dopt=s&uid=2
Structure Summary
NCBI
MMDB
Structure Summary
PubMed BLAST Structure Taxonomy OMIM Help? Cn3d
Description: Leucine(Slash)Isoleucine(Slash)Valine-Binding Protein (LIVBP).
Deposition: J.S.Sack, M.A.Saper & F.A.Quiocho, 10-Apr-89
Taxonomy: Escherichia coli
Reference: PubMed MMDB: 2778 PDB: 2LIV
View 3D Structure of Best Model with Cn3D Display NEW Get Cn3D 4.1!
Protein
3d Domains
CDs
Chain
LivK
Disclaimer | Write to the Help Desk
NCBI | NLM | NIH


VAST Summary – Netscape
http://www.ncbi.nlm.nih.gov/Structure/vast/vastsrv.cgi?sdid=6728
VAST Summary
NCBI
VAST
Structure Neighbors
PubMed BLAST Structure Taxonomy OMIM Help? Cn3D
Query: MMDB 2778, 2LIV
Description: Leucine(Slash)Isoleucine(Slash)Valine-Binding Protein (LIVBP)
View 3D Structure of All Atoms with Cn3D Display NEW Get Cn3D 4.1!
View Alignment using Hypertext for Selected VAST neighbors
List subset NR, Blast_p<10e-40 sorted by Aligned Length page 1 in Graphics
Find MMDB or PDB ids: or 3D-Domain ids:
60 out of 236 structure neighbors displayed.
2LIV
3d Dom.
CDs
Chain
LivK
Ali_Res.
2LBP 344
1EWT A 332
1DP4 C 312
1QOO B 285
1GUD A 253


VAST Summary – Netscape
http://www.ncbi.nlm.nih.gov/Structure/vast/vastsrv.cgi
VAST Summary
NCBI
VAST
Structure Neighbors
PubMed BLAST Structure Taxonomy OMIM
Query: MMDB 2778, 2LIV
Description: Leucine(Slash)Isoleucine(Slash)Valine-Binding Protein (LIVBP)
View 3D Structure of All Atoms with Cn3D Display
View Alignment using Hypertext for Selected VAST neighbors
List subset NR, Blast_p<10e-7 sorted by Vast P_value page 1 in Table
Find MMDB or PDB ids:
60 out of 146 structure neighbors displayed.
P-value ≤ 0.001
and
% Identity > 25
over at least 20 residues
PDB C D Ali. Len. SCORE P-VAL RMSD %Id Description
2LBP 344 43.5 10e-38.0 0.9 79.4 Leucine-Binding Protein (LBP)
1DP4 C 312 29.0 10e-20.5 4.1 16.3 Dimerized Hormone Binding Domain Of The Atrial Natriuretic Peptide Receptor
1Q00 B 285 32.6 10e-12.8 4.1 13.3 Amide Receptor Of The Amidase Operon Of Pseudomonas Aeruginosa (Amic) Complexed With The Negative Regulator Amir.
1Q00 B 1 222 20.2 10e-12.6 2.8 15.8 Amide Receptor Of The Amidase Operon Of Pseudomonas Aeruginosa (Amic) Complexed With The Negative Regulator Amir.


Cn3D Viewer
Rendering: Tubes
Coloring: Identity
Red – matches
Blue – mismatches
2LIV neighbors – Cn3D 4.0
2LIV neighbors – Sequence/Alignment Viewer



Worms	**Rendering**	Spacefill
Secondary Structure	**Coloring**	Charge

SWISS-MODEL

- Automated comparative protein modelling server
- Web front-end at
  http://www.expasy.org/swissmod
  Results returned by E-mail

BLAST search to find similarities in PDB *by sequence*

↓

Select templates with sequence identity > 25% and projected model size > 20 amino acids

↓

Generate models

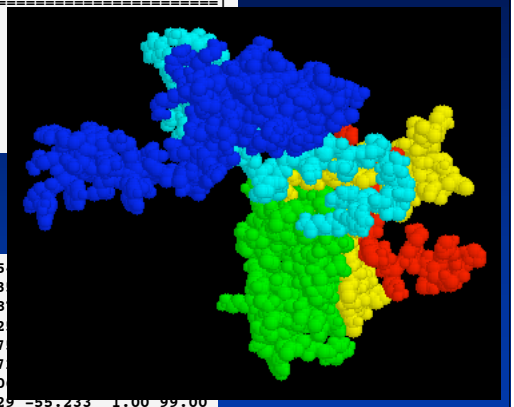
↓

Do energy minimization

↓

Generate PDB file for new protein model

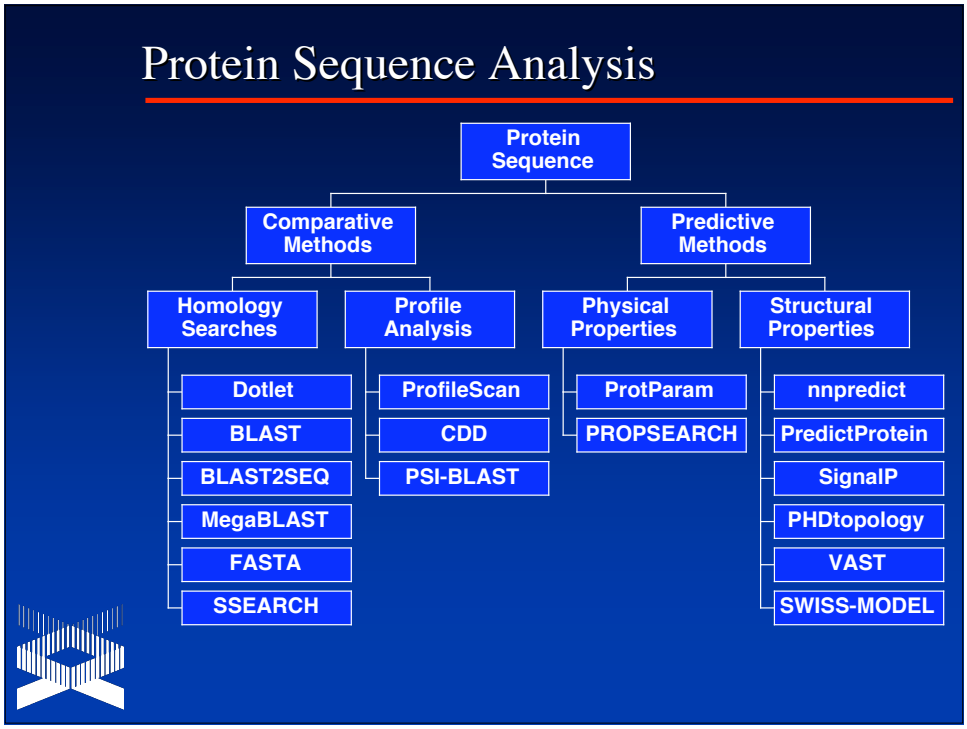
```
21DJH.pdb: 42.77 % identity
21DJG.pdb: 42.77 % identity
11DJG.pdb: 42.22 % identity
11QAS.pdb: 44.17 % identity
11QAT.pdb: 43.52 % identity
21QAT.pdb: 43.52 % identity
21QAS.pdb: 43.52 % identity

Target:          |==================================================|
21DJH.pdb        |         ---------
21DJG.pdb        |         ---------
11DJG.pdb        |         ---------
11QAS.pdb        |         ---------
11QAT.pdb        |         ---------
21QAT.pdb        |         ---------
21QAS.pdb        |         ---------
```

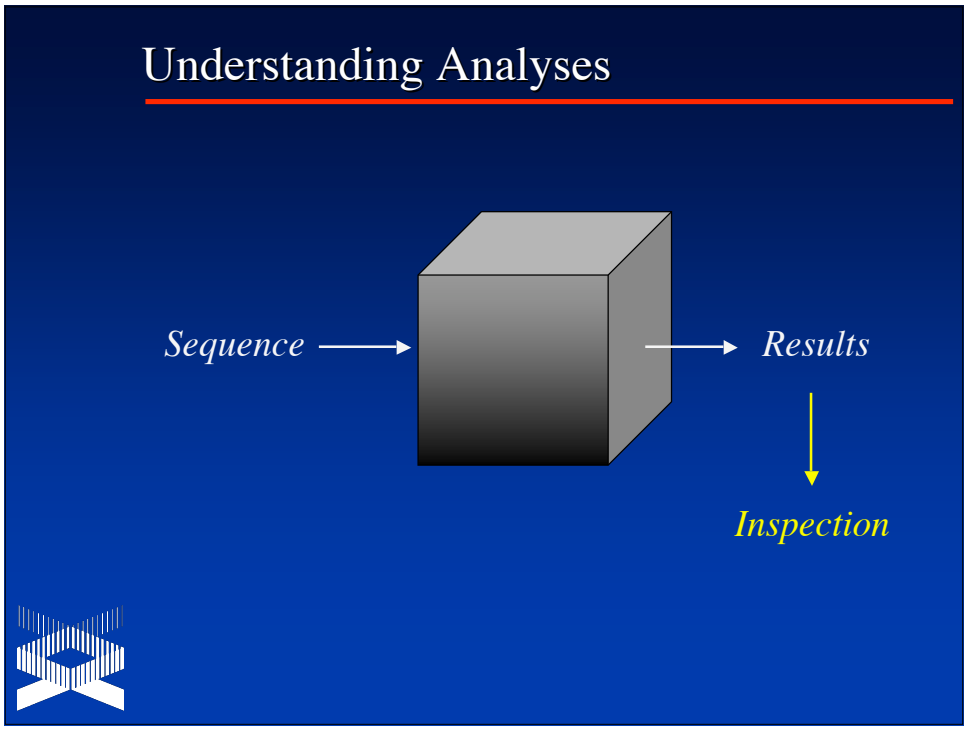


```
ATOM      1  H1  SER     1       24.219  22.95
ATOM      2  H2  SER     1       24.770  21.43
ATOM      3  N   SER     1       24.355  22.18
ATOM      4  H3  SER     1       23.466  21.92
ATOM      5  CA  SER     1       25.266  22.67
ATOM      6  CB  SER     1       24.826  24.07
ATOM      7  OG  SER     1       24.857  25.00
ATOM      8  HG  SER     1       24.717  25.929 -55.233  1.00 99.00
ATOM      9  C   SER     1       25.471  21.750 -53.751  1.00 25.00
ATOM     10  O   SER     1       25.923  22.169 -52.684  1.00 25.00
ATOM     11  N   LYS     2       25.227  20.460 -53.972  1.00 25.00
ATOM     12  H   LYS     2       24.961  20.142 -54.878  1.00 99.00
ATOM     13  CA  LYS     2       25.366  19.408 -52.943  1.00 25.00
ATOM     14  CB  LYS     2       24.003  18.772 -52.622  1.00 25.00
```

Structural Modeling Software

- 3D-JIGSAW
  http://www.bmm.icnet/uk/servers/3djigsaw
- ESyPred3D
  http://www.fundp.ac.be/urbm/bioinfo/esypred
- MODELLER
  http://www.salilab.org/modeller/modeller.html
- Protinfo
  http://protinfo.compbio.washington.edu


Protein Sequence Analysis
Protein Sequence
Comparative Methods
Predictive Methods
Homology Searches
Profile Analysis
Physical Properties
Structural Properties
Dotlet
BLAST
BLAST2SEQ
MegaBLAST
FASTA
SSEARCH
ProfileScan
CDD
PSI-BLAST
ProtParam
PROPSEARCH
nnpredict
PredictProtein
SignalP
PHDtopology
VAST
SWISS-MODEL


Understanding Analyses
Sequence
Results
Inspection

Some lessons learned by bioinformaticians – sometimes, the hard way

Simple Searches using Different Methods

- Pfam and SMART can be searched using
  - Pfam's search engine
  - SMART's search engine
  - NCBI's CDD search engine
- The methods used to search differ
  - Pfam and SMART use HMMs
  - NCBI-CDD uses RPS-BLAST
- The results obtained differ in a significant number of cases
  - CDD will miss entries representing short domains, repeats, and motifs

"Short Motif Pitfall"

- The level of sequence identity required for significant homology is much higher for smaller regions
- Two proteins may share a common domain while still being dissimilar elsewhere
- For very short motifs, homology *cannot* be inferred by sequence identity

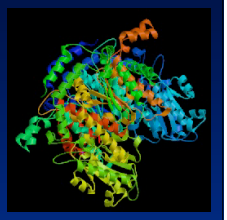
→ *short motifs may not be helpful in describing what a protein does*

Immunoglobulin Signature

- Signature defined: `[FY]-x-C-x-[VA]-x-H`
- Precision
  - Total: 480 hits in 436 sequences
  - True positives: 390 hits
  - False positives: 90 hits
  - False negatives: 23 known

Acyl-CoA dehydrogenase
Acyl-amino acid-releasing enzyme
Alpha-adaptin A
GDP-mannose 6-dehydrogenase
Membrane alanyl aminopeptidase
Phosphatidyl cytidylyl transferase
D-lactate dehydrogenase
DNA polymerase B
Hemerythyrin
Anterior-restricted homeobox protein
Mast-stem cell growth factor
Limulus clotting factor C
Arachidonate 12-lipoxygenase
Aminoadipate-semialdehyde dehydrogenase
DNA replication licensing factor
Neprin A
Cytochrome C-522
Phosphatidylinositol 3-kinase
Origin recognition complex subunit 2
Para-aminobenzoate synthase
Alpha-platelet-derived growth factor
Serine-threonine protein kinase
Photosystem II 44 kDa reaction center protein
DNA-directed RNA polymerase II (subunits)
Chloroplast 30S ribosomal protein S4
Titin

100% identity, but...



- Phosphoglucose isomerase
  catalyzes interconversion of D-glucose-6-phosphate and D-fructose-6-phosphate
- Neuroleukin
  secreted by T-cells, promotes survival of some embryonic spinal neurons and sensory nerves; B-cell maturation
- Autocrine motility factor
  tumor cell product that stimulates cancer cell migration (metastasis?)
- Differentiation and maturation mediator
  In vitro *differentiation of human myeloid leukemia HL-60 cells to terminal monocytes*

Jeffery et al., *Biochemistry* 39, *955-964, 2000*

Proteins with Multiple Functions

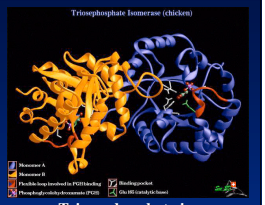
Thymidine phosphorylase	Endothelial cell growth factor
Thymidylate synthase	Translation inhibitor
birA biotin synthase	bir operon repressor
Cystic fibrosis transmembrane conductance regulator (CFTR)	Regulates other ion channels
Crystallin	Enolase Lactate dehydrogenase Heat shock protein

Does sequence similarity imply common function?

Maybe.

Structural Superfamilies: TIM Barrel

- Minimum 200 residues required for structure, with 160 residues structurally equivalent
- Structures mediate a wide variety of chemical reactions critical to biological survival
- May account for up to
  - 10% of all soluble enzymes
  - 10% of all proteins



Triose phosphate isomerase
Ribulose-phosphates
Thiamin phosphate synthase
FMN-linked oxidoreductases
NAD(P)-linked oxidoreductase
Glycosyltransferases
Metallo-dependent hydrolases
Aldolase
Enolase
Phosphenol pyruvate
Malate synthase G
RuBisCo
Xylose isomerase-like proteins
Bacterial luciferase-like proteins
Quinolinic acid phosphoribosyltransferases
Cobalamin (B12)-dependent enzymes
tRNA-guanine transglycosylase
Dihydropteroate synthetase
Uroporphyrinogen decarboxylase
Methylenetetradhydrofolate reductase
Phosphoenolpyruvate mutase

Does structural similarity imply common function?

It depends.

Annotations

- Sequence-based annotations
  - Computational predictions on raw sequence data
  - Predictions are sometimes inconsistent
- Functional annotations
  - Review of functional annotations in *Mycoplasma* (Brenner, 1999)
  - 8% of functional annotations incorrect, with many inconsistent with known *Mycoplasma* biology and metabolism

A generalized strategy for avoiding the pitfalls

Predicting Function

- Understand the limitations of the programs being used
- Identify any special features in sequence
- Identify homologous proteins
- Identify protein family members based on sequence
- Look for structural homology
- Attempt to predict the function of the protein, with appropriate cautions in mind

Identify Special Features in Sequence

- Mask sequence to reduce biologically insignificant hits
  - Non-globular regions
  - Compositionally-biased regions
  - Coiled-coil regions
- Assay for putative transmembrane regions
  - May only be significant when similarity is global
- Perform secondary structure prediction
  - Best corroboration for BLAST or other homology-based match

Identify Homologous Proteins

- Search using specialized domain databases
  - Databases include PROSITE and Pfam
  - Short motifs are of limited utility in assessing function
  - Use multiple methods (ProfileScan, SMART, CDD)
  - Look for linkage of individual domains between proteins
- Search using BLAST
  - Use appropriate weight matrix
  - Using smaller subsequences of longer proteins reduces spurious matches (*e.g.*, against kinases)
  - Use known motifs or low-complexity regions as breakpoints

Identify Homologous Proteins

Do NOT use sequence-based search methods as a "black box."

You MUST understand the methods and optimize them on a case-by-case basis

Identify Protein Family Members

- Perform iterative database searches to identify closely- and distantly-related family members
  - PSI-BLAST
  - MoST
- Construct a multiple sequence alignment
  - Look for conservation pattern between the unknown and the balance of the family to confirm presence of unique sequence features
  - Allows for assignment to family when there are few yet important sequence determinants
  - Keep in mind that sequence similarity is intransitive
    - AB ~ BC, and BC ~ CD, but AB ≠ CD

Identify Protein Family Members

- "Prediction by Analogy"
  - Catalytic site residues are almost invariably polar
  - Large aromatic residues are often found to be involved in protein-ligand interactions
  - Zinc ions are coordinated by several residue types and, often, water molecules
  - Calcium ions are often bound by acidic residues and amides, although additional interactions occur with backbone atoms
  - Mn/Mg ions are often bound by two acidic residues separated by a hydrophobic residue in nucleases and glycotransferases

— Ponting, Briefings in Bioinformatics 2, 19-29, 2001

Identify Protein Family Members

- "Prediction by Analogy"
  - Phosphate and sulfate groups are found bound to the amino-terminus of alpha-helices in approximately half of all cases
  - Distinction can be made between disulfide-rich secreted proteins and zinc fingers, since the former occur in proteins with signal peptides and never possess substitutions of cysteine for histidine, or vice-versa

— Ponting, Briefings in Bioinformatics 2, 19-29, 2001

Look for Structural Homology

- ***Structure is more conserved than sequence***
- Comparison of two known structures
  - Vector-based (NCBI VAST)
  - Energy minimization methods
- Predictive modeling methods (sequence *vs.* structure)
  - SWISS-MODEL
  - Homology model building ("threading")
  - *De novo* structure prediction

Predicting Function: Considerations

- The protein may actually have more than one function within the cell
- ***Never*** use database annotation as evidence of function...
  - when there are few homologues
  - when the homologues are not consistent
- Annotations are intransitive!
- Confirm database annotations in the literature
- Predictions are subjective

Predicting Function: Considerations

- Assure that the database hits and predictive methods based on sequence yield information that ***make biological sense***
  - Predicted motifs or features biologically correct
  - Consistency with findings at the bench
- ***Even if one is able to predict function, the prediction can indeed turn out to be incorrect – experimental proof is absolutely essential!***

Gene-Function Analysis

Genome	Complete set of genes of an organism	Systematic DNA sequencing
Transcriptome	Complete set of mRNA molecules present in a cell, tissue, or organ	Hybridization arrays SAGE High-throughput Northerns
Proteome	Complete set of protein molecules present in a cell, tissue, or organ	2D gel electrophoresis Peptide mass fingerprinting Two-hybrid analysis
Metabolome	Complete set of metabolites (low-MW intermediates) in a cell, tissue, or organ	IR spectroscopy Mass spectroscopy NMR spectroscopy