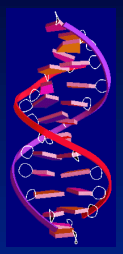
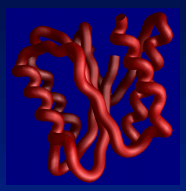
The Flow of Biotechnology Information

Gene



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

Function

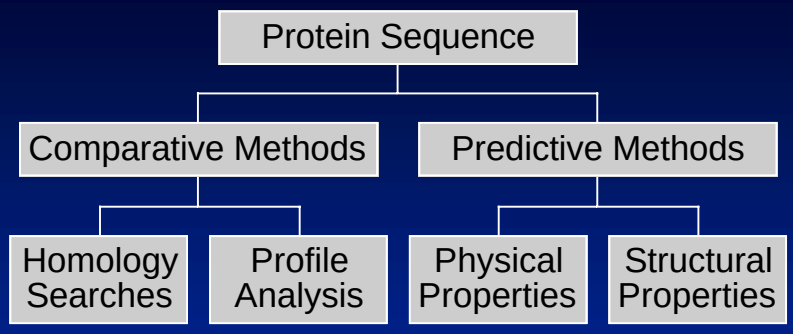


```
> Protein sequence
MKIVYWSGTGNTEKMAELIAKGIIESGKDVNTINVSDVNI
DELLNEDILILGCSAMGDEVLEESEFEPFIEEISTKISGK
KVALFGSYGWGDGKWMRDFEERMNGYGCVVVETPLIVQNE
PDEAEQDCIEFGKKIANI
```

Protein 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



- *Shared ancestry?*
- *Similar function?*
- *Domain or complete sequence?*

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

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...
```

Select ALL databases
Significant matches only

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

E-value

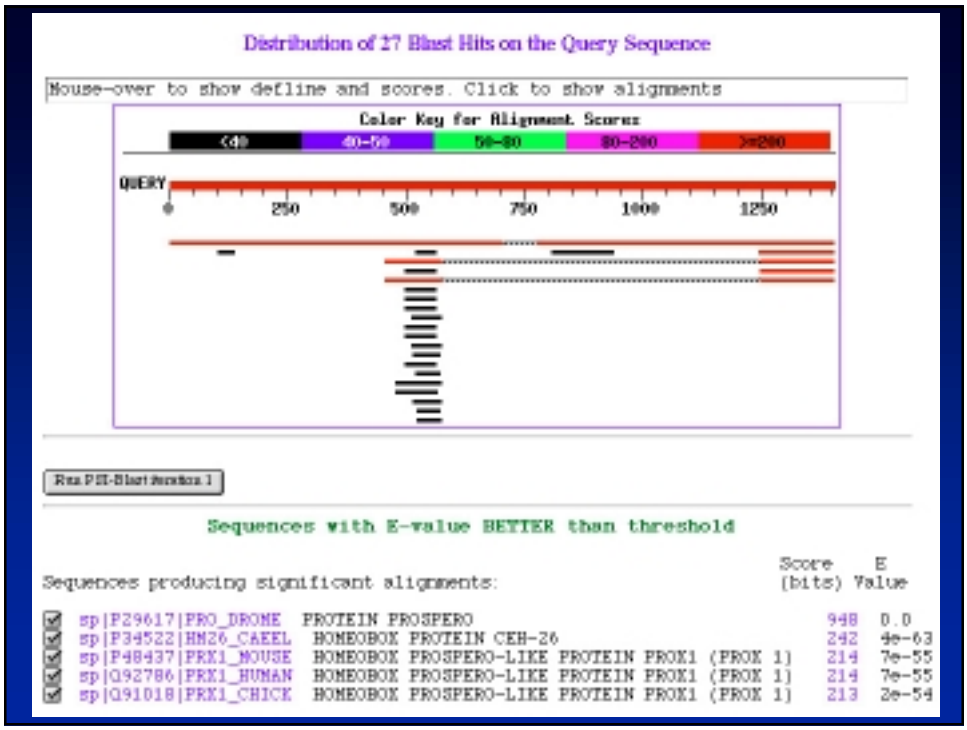
NScore	SwissProt
7.0	1.8000
8.0	0.1800
9.0	0.0180
10.0	0.0018
219.4	3e-211

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]
```

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



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*

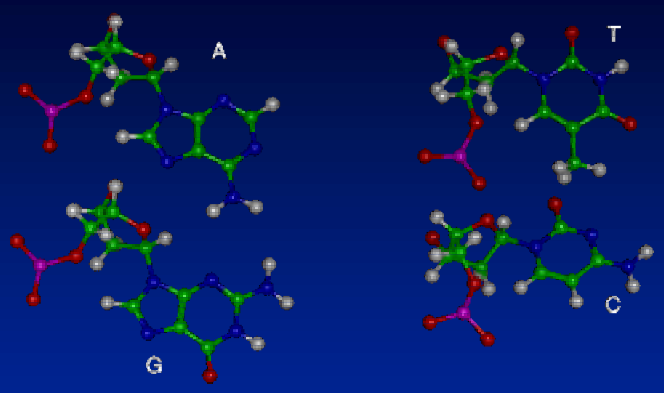
```
>Histone chk-H5  family
6 sequences are included in 2 blocks

            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

- Protein Sequence
  - Comparative Methods
    - Homology Searches
    - Profile Analysis
  - Predictive Methods
    - Physical Properties
    - Structural Properties

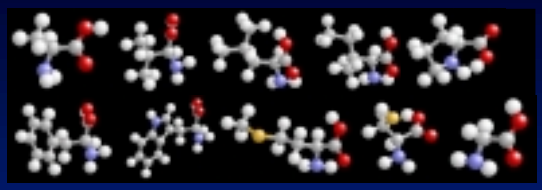
- *Composition*
- *Hydrophobicity*
- *Secondary structure*
- *Specialized structures*
- *Tertiary structure*

Information Landscape

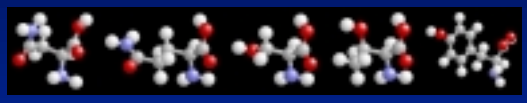


Information Landscape

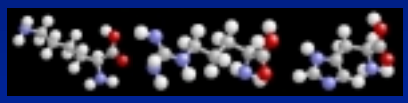
Nonpolar



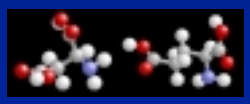
Polar *Neutral*



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

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%

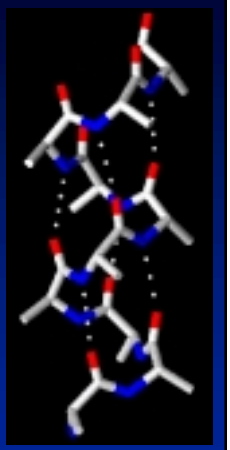
Rank	ID	DIST	LEN2	POS1	POS2	pI	DE
1	>p1;s18193	0.00	727	1	727	5.33	autoantigen NOR-90 - human
2	ubf1_human	1.36	764	1	764	5.62	NUCLEOLAR TRANSCRIPTION FACTOR 1
3	ubf1_mouse	1.40	765	1	765	5.55	NUCLEOLAR TRANSCRIPTION FACTOR 1
4	ubf1_rat	1.57	764	1	764	5.61	NUCLEOLAR TRANSCRIPTION FACTOR 1
5	ubf1_xenla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

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

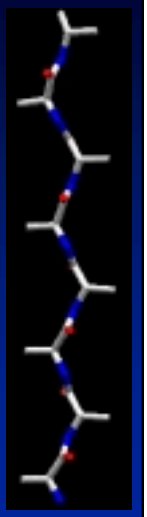
Alpha-helix

- 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



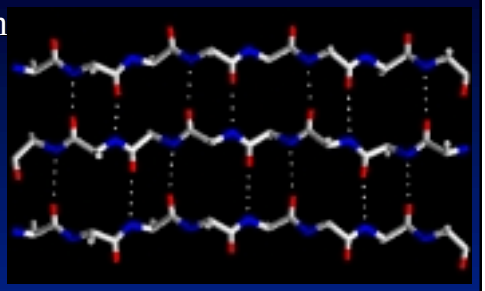
Beta-strand

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



Beta-sheet

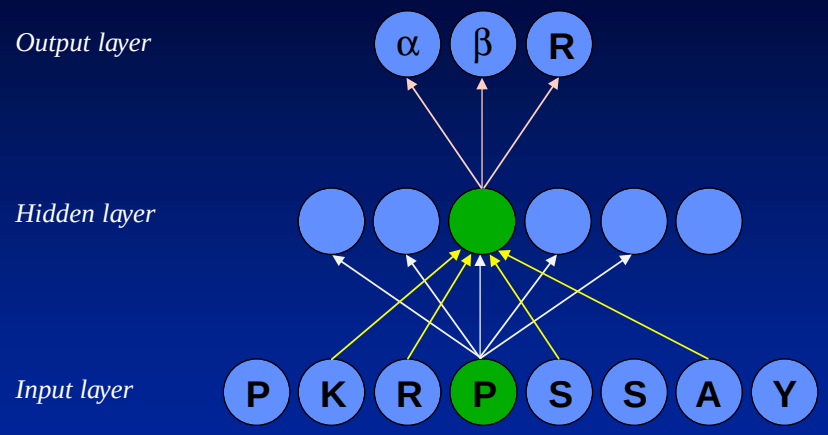
- Stabilization through hydrogen bonding
- Parallel or antiparallel
- Variant: beta-turn



Folding Classes

α	β	α+β	α/β
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

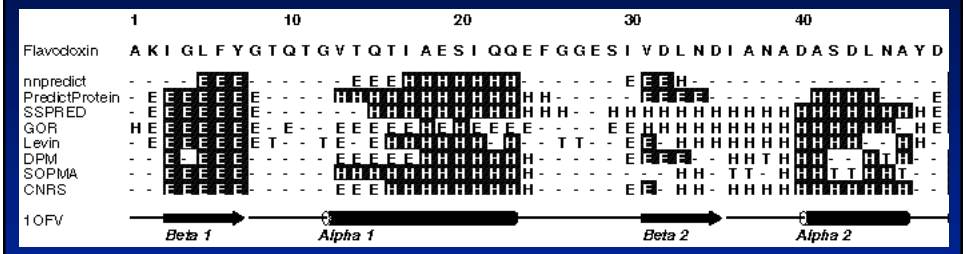
```
Joe Buzzcut
National Human Genome Research Institute, NIH
buzzcut@nhgri.nih.gov
# flavodoxin - Anacystis nidulans
AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADASDLNAYDYLIIGCPTWNVGELQSDWEGIY
DDLDSVNFQGKKVAYFGAGDQVGYSDNFQDAMGILEEKISSLGSQTVGYWPIEGYDFNESKAVRNNQFVG
LAIDEDNQPDLTKNRIKTWVSQLKSEFGL
```

Secondary structure

```
          ....,....1....,....2....,....3....,....4....,....5....,....6
AA       |AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADASDLNAYDYLIIGCPTWNVG|
PHD sec  | EEEEEEE    HHHHHHHHHHHHH     EEEEE HHH HHHH    EEEEE       |
Rel sec  |938999736982489999999999767982443213241278631241999861547765|
Detail:
prH sec  |0000000000146899999999998210000111125653887643210000011111111|
prE sec  |0589988520000000000000000000036655421000000000148998741200002|
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

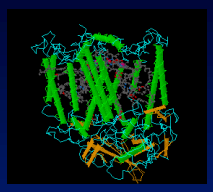
```
>sp|P05019|IGFB_HUMAN INSULIN-LIKE GROWTH FACTOR IB PRECURSOR
MGKISSLPTQLFKCCFCDFLKVKMHTMSSSHLFYLALCLLTFTSSATAGPETLCGAELVDALQFVCGDRG
```

N-terminal end only
Eukaryotic set

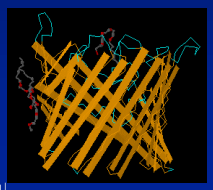
```
************************* SignalP predictions *************************
Using networks trained on euk data

>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

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 |                      HHHHHHHHHHHHHH              HHHHHHHHHH|
        Rel htm |36889999999999999864110466777765543125777888777762146778888 8|
 detail:        |                                                            |
        prH htm |31000000000000000012445788888887776532110000111135788899999|
        prL htm |68999999999999999987554211111122234678889999888864211100000|
                                          .
                                          .
                                          .
        PHDThtm |iiiiiiiiiiiiiiiiiiTTTTTTTTTTTTTTTTTTooooooooooooooooTTTTTTTT|
```

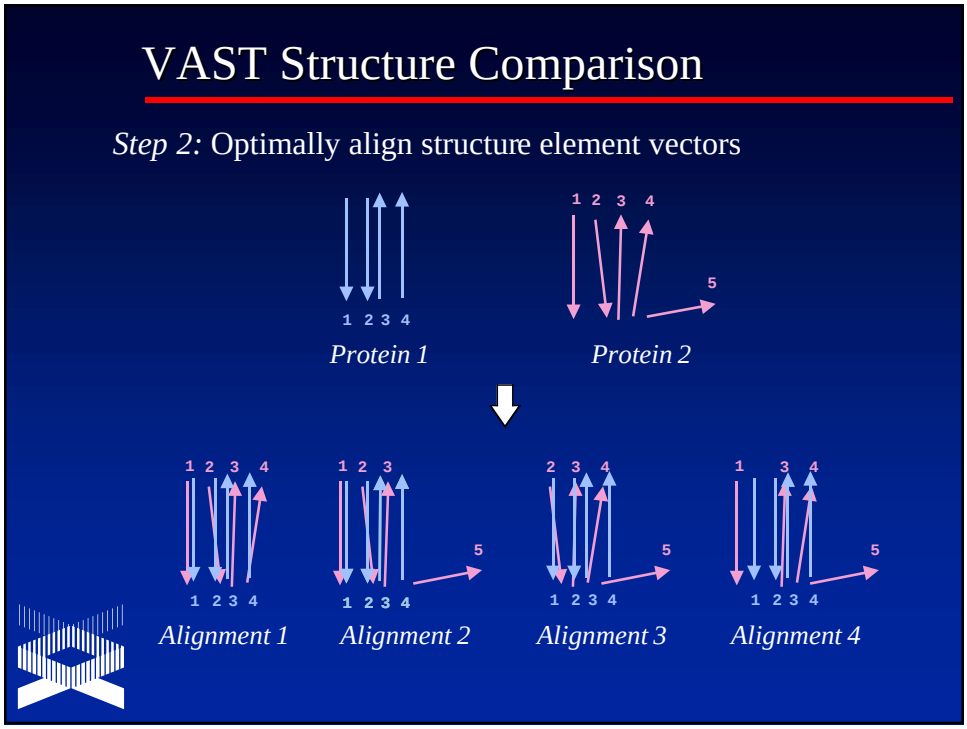
Predicting Tertiary Structure

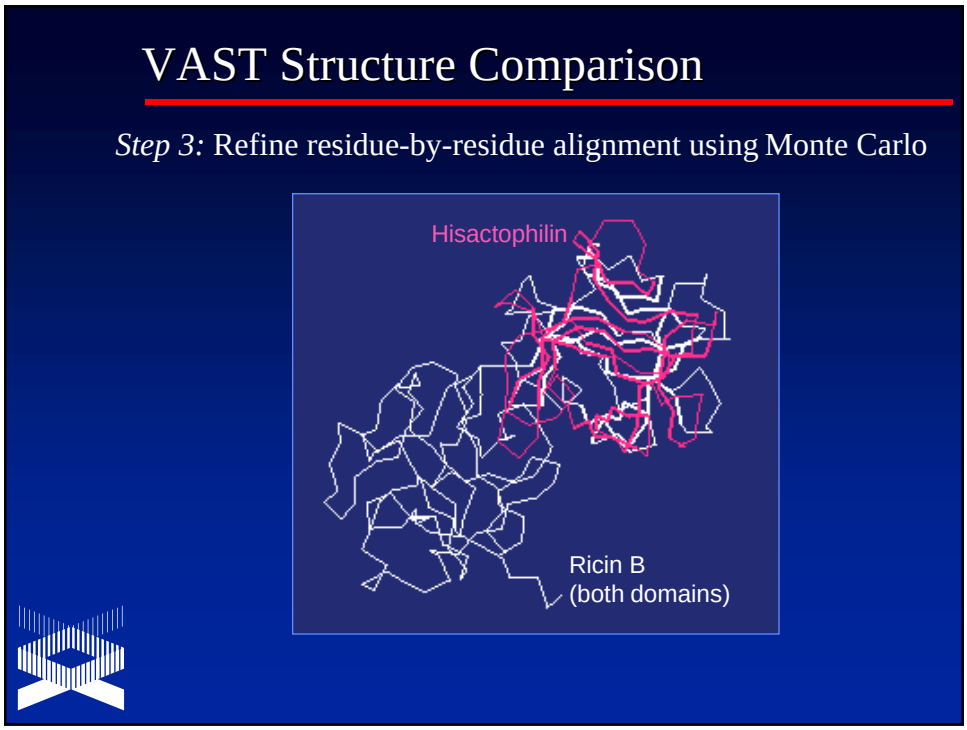
- Sequence specifies conformation, *but* conformation does *not* specify sequence
- Structure is conserved to a much greater extent than sequence
  - Limited number of protein folds
- Similarities between proteins may not necessarily be detected through "traditional" methods

VAST Structure Comparison

Step 1: Construct vectors for secondary structure elements

Ricin Chain B


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 using Monte Carlo
Hisactophilin
Ricin B
(both domains)

MMDB Id: 2778 PDB Id: 2LIV

Protein Chains: (single chain)
MEDLINE: PubMed
Taxonomy: Escherichia coli

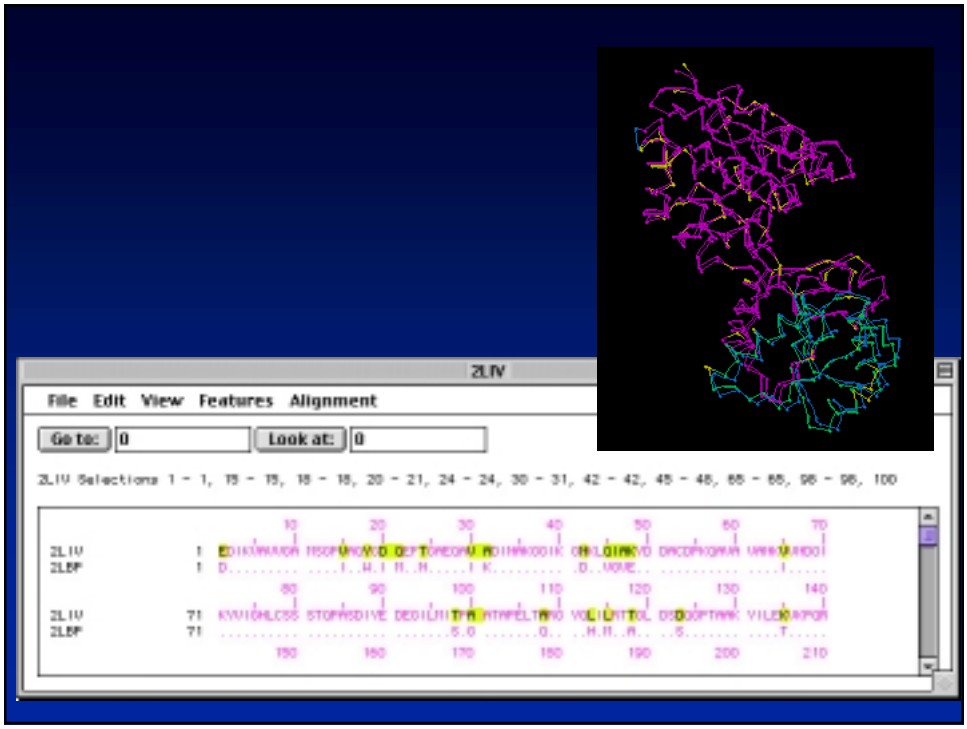
PDB Authors: J.S.Sack, M.A.Saper & F.A.Quiocho
PDB Deposition: 10-Apr-89
PDB Class: Periplasmic Binding Protein
PDB Compound: Leucine(Slash)Isoleucine(Slash)Valine-Binding Protein (LIVBP)

Sequence Neighbors: (single chain)
Structure Neighbors: (single chain), 1, 2

View / Save Structure NEW *Get Cn3D 2.0 Now!*

Options: Launch Viewer; See File; Save File
Viewer: Cn3D v2.0 (asn.1); Cn3D v1.0 (asn.1); Mage; RasMol (PDB)
Complexity: Cn3D Subset; Virtual Bond Model; All Atom Model; Up to 5 Models; Up to 10 Models; All Models

	PDB	C	D	SCO	P-VAL	RMSD	NRES	%Id	Description
☐	2LBP		1	22.8	10e-18.7	0.9	263	76.4	Leucine-Binding Protein (LBP)
☐	1PEA		1	20.7	10e-14.3	3.1	217	14.3	Amide ReceptorNEGATIVE REGULATOR OF THE AMIDASE OPERON OF Pseudomonas Aeruginosa (Amic) Complexed With Acetamide
☐	1BMT	A	2	15.1	10e-7.4	2.9	120	8.3	Methionine Synthase (B12-Binding Domains) (E.C.2.1.1.13)
☐	1DHR			17.1	10e-7.0	4.3	111	10.8	Dihydropteridine Reductase (Dhpr) (E.C.1.6.99.10) Complex With Nadh
☐	8ABP		1	14.9	10e-7.0	3.2	125	10.4	L-Arabinose-Binding Protein (Mutant With Met 108 Replaced By Leu) (M108L) Complex With D-Galactose
☐	1SCU	A	2	14.5	10e-7.0	2.5	101	10.9	Succinyl-Coa Synthetase (Succinate-Coa Ligase) (Adp-Forming) (E.C.6.2.1.5)
☐	2LBP		2	14.4	10e-6.7	2.8	110	10.0	Leucine-Binding Protein (LBP)
☐	2CUT			13.5	10e-6.1	3.0	114	7.9	Cutinase (E.C.3.1.1.-) Complexed With The Inhibitor Diethyl Para-Nitrophenyl Phosphate



SWISS-MODEL

- Automated comparative protein modelling server
- *http://www.expasy.ch/swissmod/SWISS-MODEL.html*
  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

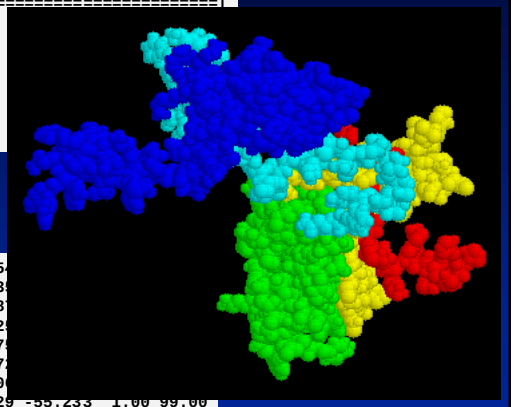
↓

Do energy minimization

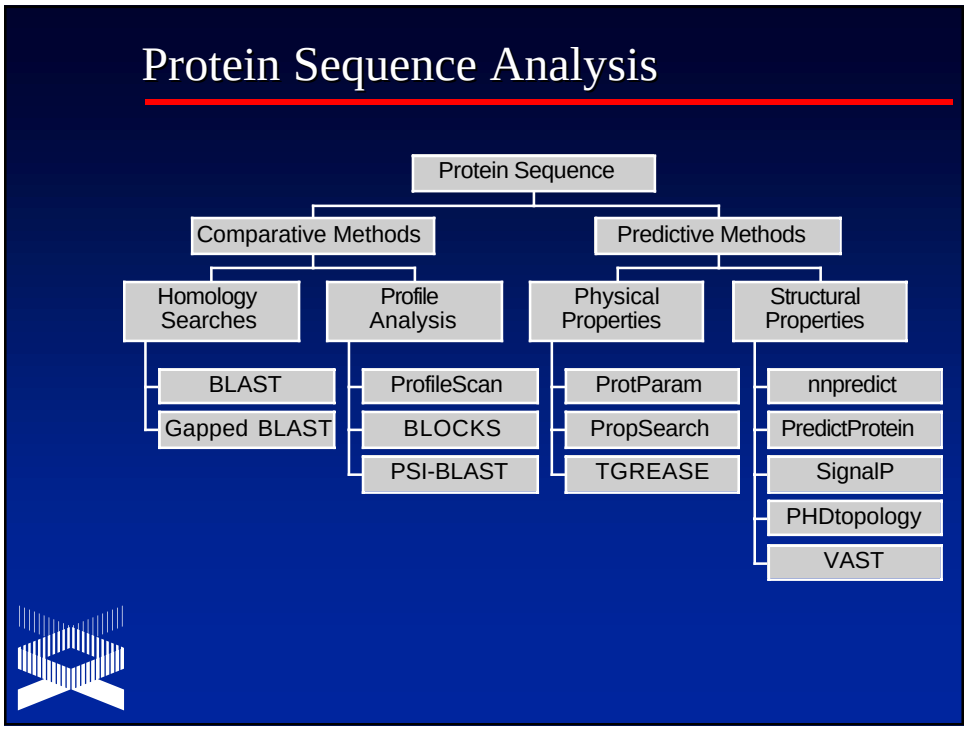
↓

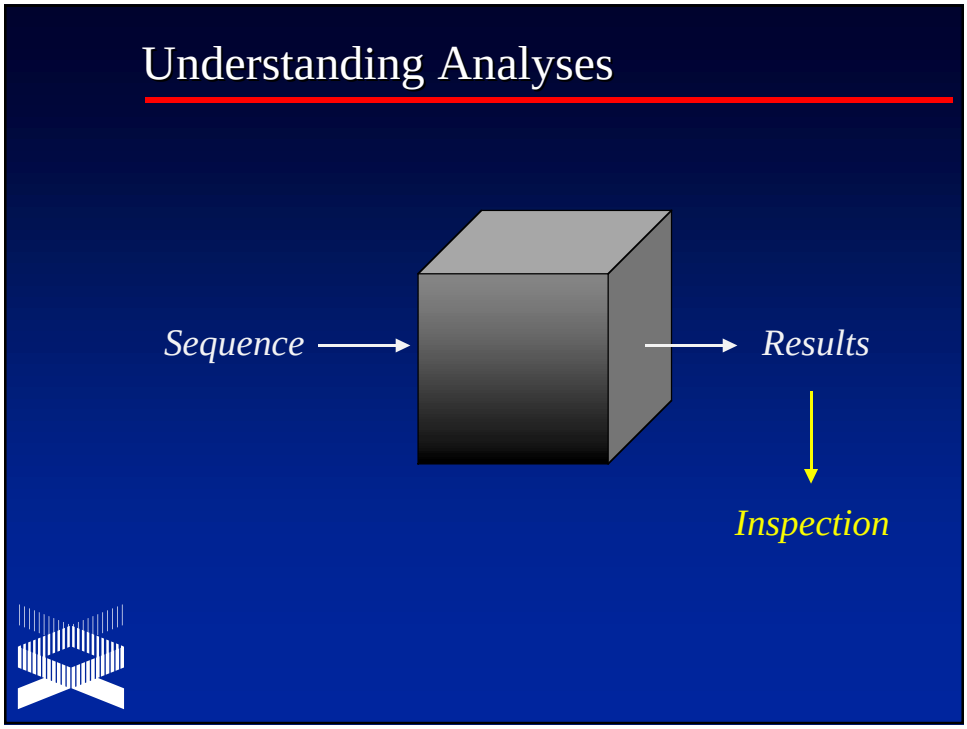
Generate PDB file for new protein model

```
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  22.954
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
```





Some lessons learned by bioinformaticians – sometimes, the hard way

"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

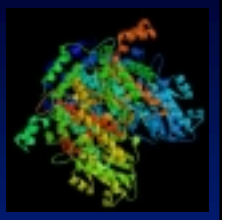
→ *short motifs may not be helpful in describing what a protein does*

Immunoglobulin Signature

- Signature defined: `[FY]-x-C-x-[VA]-x-H`
- Precision
  - Total: 456 hits in 412 sequences
  - True positives: 385 hits in 341 sequences
  - False positives: 71 hits in 71 sequences

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

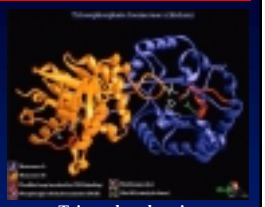
Thymidine phosphorylase	Endothelial cell growth factor
Thymidylate synthase	Translation inhibitor
birA biotin synthase	bir operon repressor
Cystic fibrosis transmembrane conductance regulator (CFTR)	Regulates other ion channels
Crystallin	Enolase Lactate dehydrogenase Heat shock protein

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.

Predicting Function

- 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
- 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
- ***Do not use sequence-based search methods as a "black box" – the user 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 $\not\sim$ CD

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

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