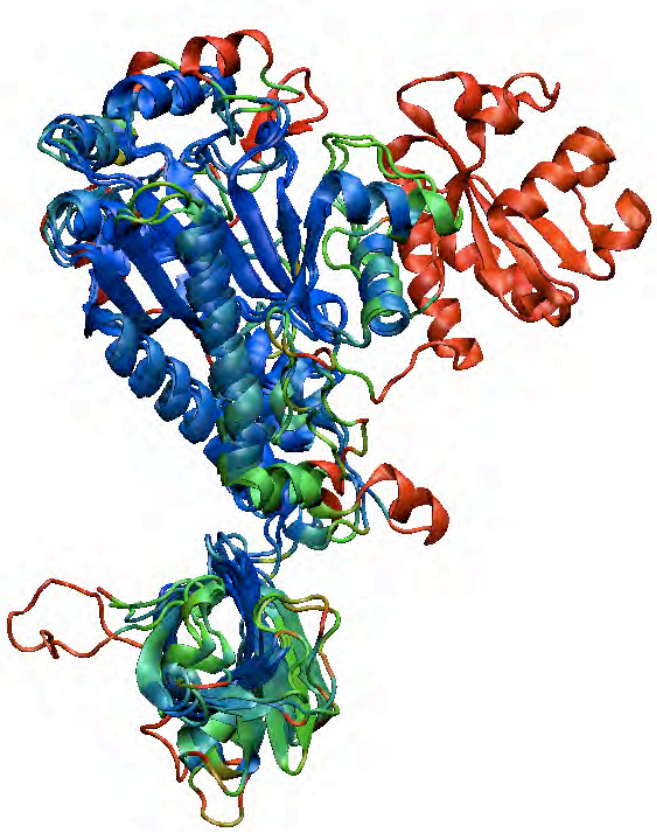
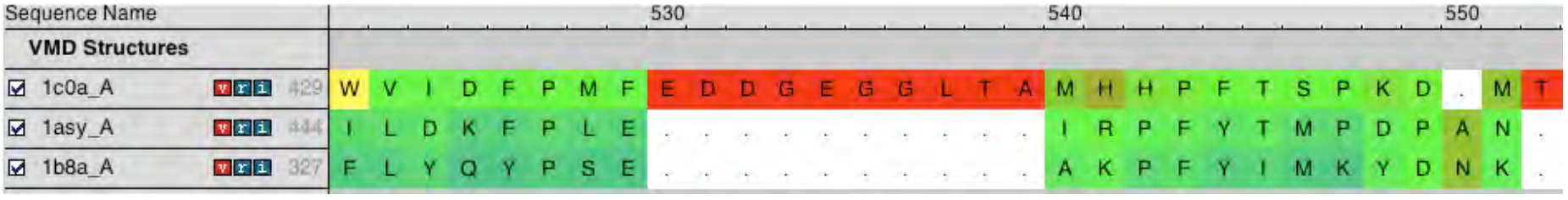
Sequence and Structure Alignment

Z. Luthey-Schulten, UIUC

Frankfurt, Germany, 2006

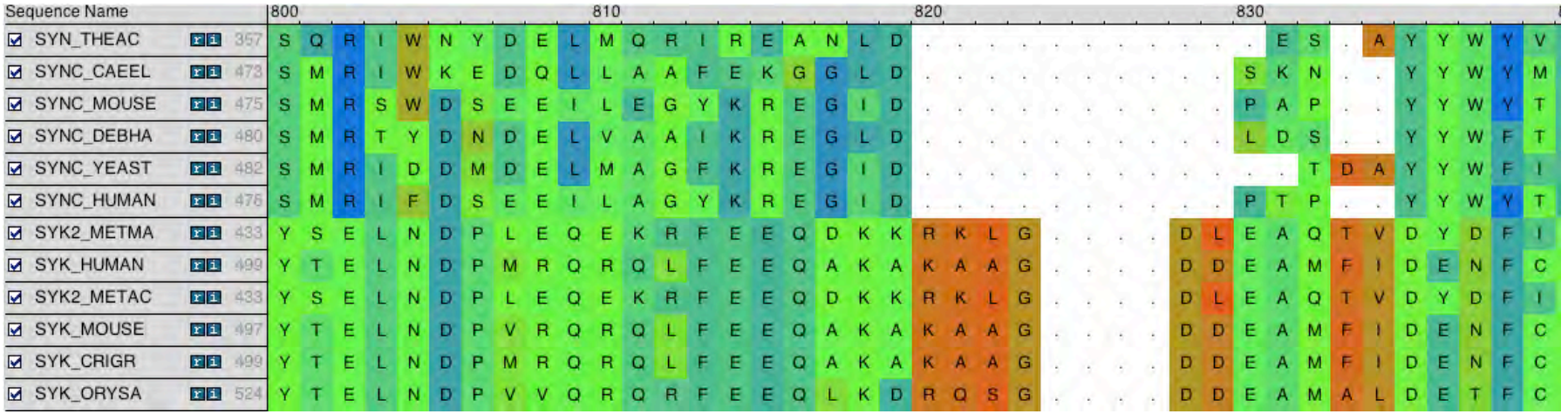


VMD 1.8.4



Why Look at More Than One Sequence?

1. Multiple Sequence Alignment shows patterns of conservation



2. What and how many sequences should be included?

3. Where do I find the sequences and structures for MS alignment?

4. How to generate pairwise and multiple sequence alignments?

Sequence-Sequence Alignment

- Smith-Watermann
- Needleman-Wunsch

Seq. 1: $a_1\ a_2\ a_3$ - - $a_4\ a_5 \ldots a_n$

Seq. 2: c_1 - $c_2\ c_3\ c_4\ c_5$ - $\ldots c_m$

Sequence-Structure Alignment

- Threading
- Hidden Markov, Clustal

Structure-Structure Alignment

- STAMP - Barton and Russell
- CE - Bourne et al.

Sequence Database Searches

- Blast and Psi-Blast

Sequence-Sequence Alignment

- Smith-Watermann
- Needleman-Wunsch

Profile 1: $A_1\ A_2\ A_3$ - - $A_4\ A_5 \ldots A_n$

Profile 2: C_1 - $C_2\ C_3\ C_4\ C_5$ - $\ldots C_m$

Sequence-Structure Alignment

- Threading
- Hidden Markov, Clustal

Structure-Structure Alignment

- STAMP - Barton and Russell
- CE - Bourne et al.

SCOP, Astral

PDB

Sequence Database Searches

- Blast and Psi-Blast NCBI Swiss Prot

Swiss-Prot

Protein knowledgebase

TrEMBL

Computer-annotated supplement to Swiss-Prot

UniProt
the universal protein resource

The UniProt Knowledgebase consists of:

- **Swiss-Prot**; a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases [More details / References / Linking to Swiss-Prot / User manual / Recent changes / Commercial users / Disclaimer].
- **TrEMBL**; a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot.

These databases are developed by the Swiss-Prot groups at SIB and at EBI.

UniProt Release 3.2 consists of:
Swiss-Prot Release 45.2 of 23-Nov-2004: 164201 entries (More statistics)
TrEMBL Release 28.2 of 23-Nov-2004: 1503829 entries (More statistics)

> Swiss-Prot headlines
Major update of C.elegans entries (Read more...)

Search in UniProt Knowledgebase (Swiss-Prot and TrEMBL) for: aspartyl synthetase

UniProtKB/Swiss-Prot Release 49.2 of 07-Mar-2006
UniProtKB/TrEMBL Release 32.2 of 07-Mar-2006

- Number of sequences found in UniProt Knowledgebase (Swiss-Prot$_{(201)}$ and TrEMBL)$_{(259)}$: **460**
- Note that the selected sequences can be saved to a file to be later retrieved; to do so, go to the bottom of this page.
- For more directed searches, you can use the Sequence Retrieval System SRS.

Search in UniProtKB/Swiss-Prot: There are matches to 201 out of 211104 entries

SYD1_STRMU (**Q8DSG3**)
Aspartyl-tRNA synthetase 1 (EC 6.1.1.12) (Aspartate--tRNA ligase 1) (AspRS 1). {GENE: Name=aspS1; OrderedLocusNames=SMU.1822} - Streptococcus mutans

SYD2_STRMU (**Q8DRV9**)
Aspartyl-tRNA synthetase 2 (EC 6.1.1.12) (Aspartate--tRNA ligase 2) (AspRS 2). {GENE: Name=aspS2; OrderedLocusNames=SMU.2101} - Streptococcus mutans

SYDC_YEAST (**P04802**)
Aspartyl-tRNA synthetase, cytoplasmic (EC 6.1.1.12) (Aspartate--tRNA ligase) (AspRS). {GENE: Name=DPS1; Synonyms=APS, APS1; OrderedLocusNames=YLL018C; ORFNames=L1295} - Saccharomyces cerevisiae (Baker's yeast)

SYDM_YEAST (**P15179**)
Aspartyl-tRNA synthetase, mitochondrial (EC 6.1.1.12) (Aspartate--tRNA ligase) (AspRS). {GENE: Name=MSD1; OrderedLocusNames=YPL104W; ORFNames=LPG5W} - Saccharomyces cerevisiae (Baker's yeast)

SYD_ACIAD (**Q6FEH6**)
Aspartyl-tRNA synthetase (EC 6.1.1.12) (Aspartate--tRNA ligase) (AspRS). {GENE: Name=aspS; OrderedLocusNames=ACIAD0609} - Acinetobacter sp. (strain ADP1)

SYD_AERPE (**Q9Y9U7**)
Aspartyl-tRNA synthetase (EC 6.1.1.12) (Aspartate--tRNA ligase) (AspRS). {GENE: Name=aspS; OrderedLocusNames=APE2192} - Aeropyrum pernix

UniProtKB/Swiss-Prot entry P04802

Entry information	
Entry name	**SYDC_YEAST**
Primary accession number	**P04802**
Secondary accession numbers	None
Integrated into Swiss-Prot on	August 13, 1987
Sequence was last modified on	October 1, 1994 (Sequence version 2)
Annotations were last modified on	March 7, 2006 (Entry version 72)
Name and origin of the protein	
Protein name	**Aspartyl-tRNA synthetase, cytoplasmic**
Synonyms	**EC 6.1.1.12** **Aspartate--tRNA ligase** **AspRS**
Gene name	**Name: DPS1** Synonyms: APS, APS1 OrderedLocusNames: YLL018C ORFNames: L1295
From	Saccharomyces cerevisiae (Baker's yeast) [TaxID: 4932]
Taxonomy	Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces.

References

[1] PROTEIN SEQUENCE.
PubMed=3902099 [NCBI, ExPASy, EBI, Israel, Japan]
Amiri I., Mejdoub H., Hounwanou N., Boulanger Y., Reinbolt J.;
"The complete amino acid sequence of cytoplasmic aspartyl-tRNA synthetase from Saccharomyces cerevisiae.";

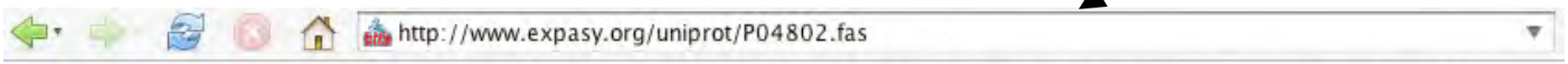
Cross-references	
Sequence databases	
EMBL	X03606; CAA27269.1; -; Genomic_DNA.[EMBL / GenBank / DDBJ] [CoDingSequence] X06665; CAA29865.1; -; Genomic_DNA.[EMBL / GenBank / DDBJ] [CoDingSequence] X97560; CAA66172.1; -; Genomic_DNA.[EMBL / GenBank / DDBJ] [CoDingSequence] Z73123; CAA97464.1; -; Genomic_DNA.[EMBL / GenBank / DDBJ] [CoDingSequence] Z73122; CAA97463.1; -; Genomic_DNA.[EMBL / GenBank / DDBJ] [CoDingSequence] X91488; CAA62772.1; -; Genomic_DNA.[EMBL / GenBank / DDBJ] [CoDingSequence]
PIR	A23508; SYBYDC.
3D structure databases	
PDB	1ASY; X-ray; A/B=67-556.[ExPASy / RCSB / EBI] 1ASZ; X-ray; A/B=67-556.[ExPASy / RCSB / EBI] 1EOV; X-ray; A=70-556. [ExPASy / RCSB / EBI] Detailed list of linked structures.
ModBase	P04802.
Protein-protein interaction databases	
IntAct	P04802; -.
DIP	P04802.
Protein family/group databases	
GermOnline	142013; -.
2D gel databases	
SWISS-2DPAGE	Get region on 2D PAGE.
Organism-specific gene databases	
SGD	S000003941; DPS1.
Yeast-GFP	YLL018C.
Ontologies	
GO	GO:0004815; Molecular function: aspartate-tRNA ligase activity *(inferred from direct assay)*. GO:0042802; Molecular function: identical protein binding *(inferred from physical interaction)*. GO:0003723; Molecular function: RNA binding *(inferred from direct assay)*. QuickGo view.
Family and domain databases	
InterPro	IPR004523; AspS_arch. IPR012340; OB_NA_bd_sub. IPR004365; OB_tRNA_NA_bd. IPR004364; tRNA-synt_2. IPR002312; tRNA-synt_asp. IPR006195; tRNA_ligase_II. Graphical view of domain structure.
Pfam	PF00152; tRNA-synt_2; 1. PF01336; tRNA_anti; 1. Pfam graphical view of domain structure.

Sequence information

Length: **556 AA** [This is the length of the unprocessed precursor] | Molecular weight: **63384 Da** [This is the MW of the unprocessed precursor] | CRC64: **D2EE179B24F25297** [This is a checksum on the sequence]

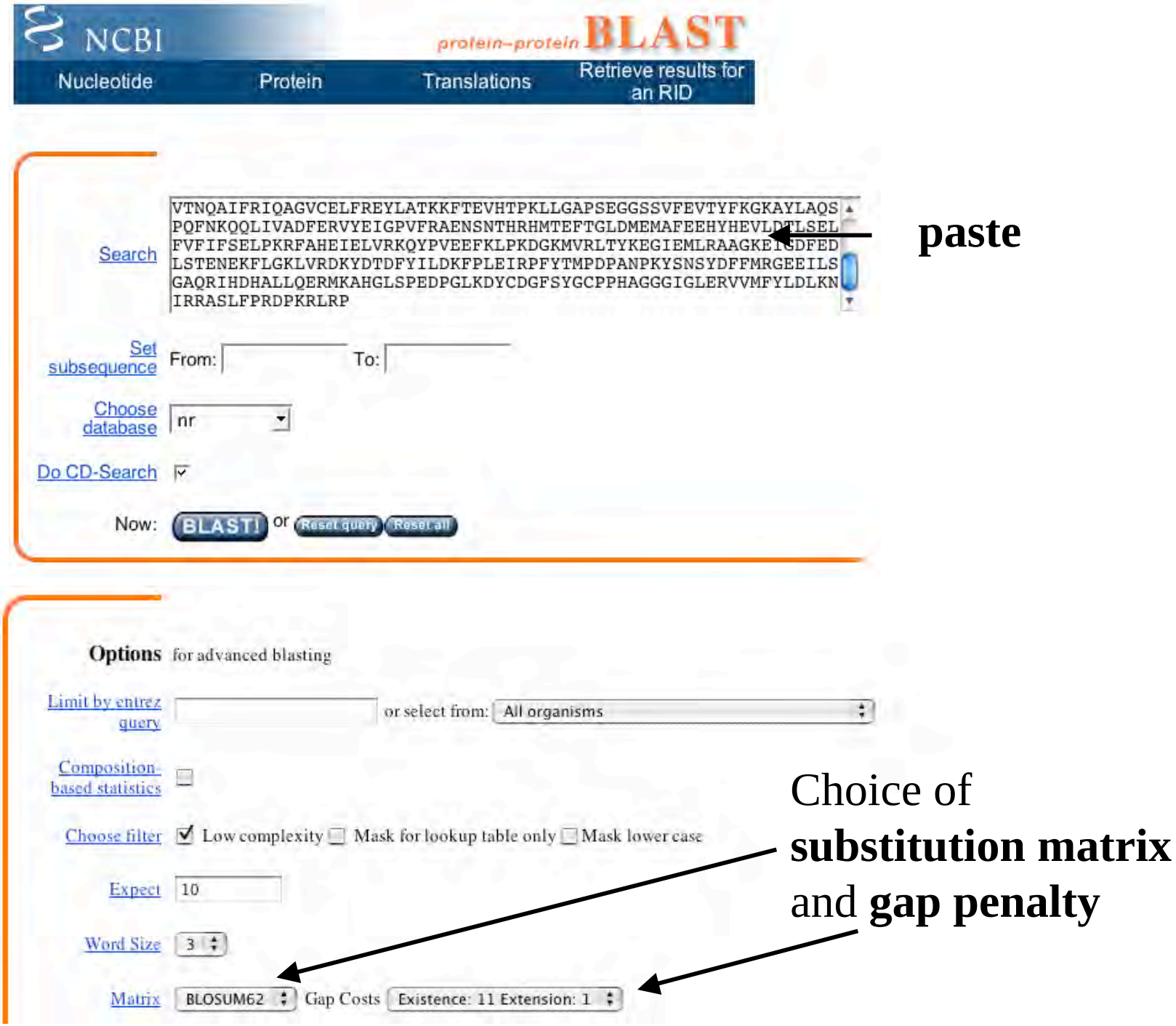
```
        10         20         30         40         50         60
SQDENIVKAV EESAEPAQVI LGEDGKPLSK KALKKLQKEQ EKQRKKEERA LQLEAEREAR

        70         80         90        100        110        120
EKKAAAEDTA KDNYGKLPLI QSRDSDRTGQ KRVKFVDLDE AKDSDKEVLF RARVHNTRQQ

       130        140        150        160        170        180
GATLAFLTLR QQASLIQGLV KANKEGTISK NMVKWAGSLN LESIVLVRGI VKKVDEPIKS

       190        200        210        220        230        240
ATVQNLEIHI TKIYTISETP EALPILLEDA SRSEAEAEAA GLPVVNLDTR LDYRVIDLRT

       250        260        270        280        290        300
VTNQAIFRIQ AGVCELFREY LATKKFTEVH TPKLLGAPSE GGSSVFEVTY FKGKAYLAQS

       310        320        330        340        350        360
PQFNKQQLIV ADFERVYEIG PVFRAENSNT HRHMTEFTGL DMEMAFEEHY HEVLDTLSEL

       370        380        390        400        410        420
FVFIFSELPK RFAHEIELVR KQYPVEEFKL PKDGKMVRLT YKEGIEMLRA AGKEIGDFED

       430        440        450        460        470        480
LSTENEKFLG KLVRDKYDTD FYILDKFPLE IRPFYTMPDP ANPKYSNSYD FFMRGEEILS

       490        500        510        520        530        540
GAQRIHDHAL LQERMKAHGL SPEDPGLKDY CDGFSYGCPP HAGGGIGLER VVMFYLDLKN

       550
IRRASLFPRD PKRLRP
```

P04802 in FASTA format




```
>sp|P04802|SYDC_YEAST Aspartyl-tRNA synthetase, cytoplasmic (EC 6.1.1.12) (Aspartate--tRNA lig
SQDENIVKAVEESAEPAQVILGEDGKPLSKKALKKLQKEQEKQRKKEERALQLEAEREAR
EKKAAAEDTAKDNYGKLPLIQSRDSDRTGQKRVKFVDLDEAKDSDKEVLFRARVHNTRQQ
GATLAFLTLRQQASLIQGLVKANKEGTISKNMVKWAGSLNLESIVLVRGIVKKVDEPIKS
ATVQNLEIHITKIYTISETPEALPILLEDASRSEAEAEAAGLPVVNLDTRLDYRVIDLRT
VTNQAIFRIQAGVCELFREYLATKKFTEVHTPKLLGAPSEGGSSVFEVTYFKGKAYLAQS
PQFNKQQLIVADFERVYEIGPVFRAENSNTHRHMTEFTGLDMEMAFEEHYHEVLDTLSEL
FVFIFSELPKRFAHEIELVRKQYPVEEFKLPKDGKMVRLTYKEGIEMLRAAGKEIGDFED
LSTENEKFLGKLVRDKYDTDFYILDKFPLEIRPFYTMPDPANPKYSNSYDFFMRGEEILS
GAQRIHDHALLQERMKAHGLSPEDPGLKDYCDGFSYGCPPHAGGGIGLERVVMFYLDLKN
IRRASLFPRDPKRLRP
```

cut


NCBI
protein-protein BLAST
Nucleotide
Protein
Translations
Retrieve results for an RID
Search
VTNQAIFRIQAGVCELFREYLATKKFTEVHTPKLLGAPSEGGSSVFEVTYFKGKAYLAQS
PQFNKQQLIVADFERVYEIGPVFRAENSNTHRHMTEFTGLDMEMAFEEHYHEVLDTLSEL
FVFIFSELPKRFAHEIELVRKQYPVEEFKLPKDGKMVRLTYKEGIEMLRAAGKEIGDFED
LSTENEKFLGKLVRDKYDTDFYILDKFPLEIRPFYTMPDPANPKYSNSYDFFMRGEEILS
GAQRIHDHALLQERMKAHGLSPEDPGLKDYCDGFSYGCPPHAGGGIGLERVVMFYLDLKN
IRRASLFPRDPKRLRP
paste
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:
All organisms
Composition-based statistics
Choose filter
Low complexity
Mask for lookup table only
Mask lower case
Expect
10
Word Size
3
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
Choice of
substitution matrix
and gap penalty

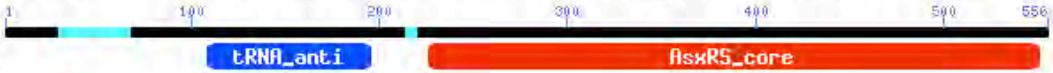
NCBI formatting BLAST

Nucleotide | Protein | Translations | Retrieve results for an RID

Your request has been successfully submitted and put into the Blast Queue.

Query = (556 letters)

Putative conserved domains have been detected, click on the image below for detailed results.



The request ID is 1142849773-4275-20262953646.BLASTQ1

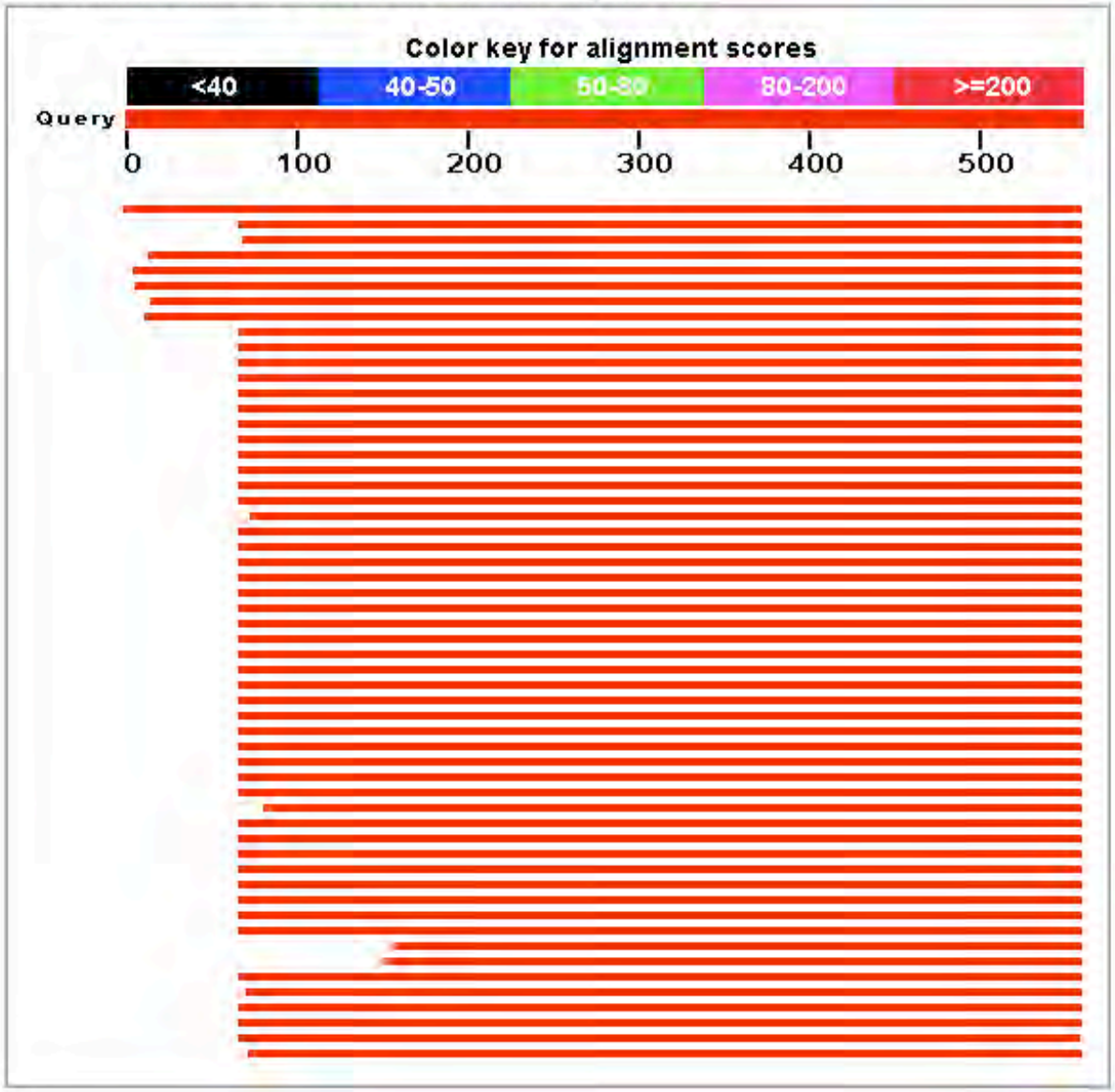
Format! or Reset all

The results are estimated to be ready in 13 seconds but may be done sooner.

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.

Distribution of 102 Blast Hits on the Query Sequence

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



Sequences producing significant alignments:	Score (Bits)	E Value
gi|6323011|ref|NP_013083.1| Cytoplasmic aspartyl-tRNA synthet...	992	0.0
gi|1064988|pdb|1ASZ|B Chain B, Aspartyl Trna Synthetase (Aspr...	961	0.0
gi|10835611|pdb|1EOV|A Chain A, Free Aspartyl-Trna Synthetase...	959	0.0
gi|49524995|emb|CAG58601.1| unnamed protein product [Candida ...	831	0.0
gi|50306525|ref|XP_453236.1| unnamed protein product [Kluyver...	813	0.0
gi|44982618|gb|AAS51881.1| ADL039Cp [Ashbya gossypii ATCC 108...	788	0.0
gi|50409277|ref|XP_456856.1| hypothetical protein DEHA0A12551...	726	0.0
gi|68485621|ref|XP_713293.1| putative aspartate-tRNA syntheta...	719	0.0
gi|50551341|ref|XP_503144.1| hypothetical protein [Yarrowia l...	693	0.0
gi|71020373|ref|XP_760417.1| hypothetical protein UM04270.1 [...	565	3e-159
gi|3618213|emb|CAA20876.1| SPCC1223.07c [Schizosaccharomyces ...	552	1e-155
gi|57226309|gb|AAW42769.1| aspartate-tRNA ligase, putative [C...	539	2e-151
gi|68365838|ref|XP_686219.1| PREDICTED: similar to Aspartyl-t...	537	4e-151
gi|27503265|gb|AAH42227.1| Dars-prov protein [Xenopus laevis]	531	3e-149
gi|73984263|ref|XP_848666.1| PREDICTED: similar to Aspartyl-t...	530	4e-149
gi|53133416|emb|CAG32037.1| hypothetical protein [Gallus gallus]	530	6e-149
gi|49522592|gb|AAH75373.1| Aspartyl-tRNA synthetase [Xenopus tro	530	8e-149
gi|39794467|gb|AAH64273.1| Aspartyl-tRNA synthetase [Xenopus ...	530	9e-149
gi|68365842|ref|XP_708309.1| PREDICTED: similar to Aspartyl-t...	529	1e-148
gi|55741590|ref|NP_001006528.1| aspartyl-tRNA synthetase [Gal...	527	4e-148
gi|49119135|gb|AAH72839.1| MGC80207 protein [Xenopus laevis]	527	5e-148
gi|74267650|gb|AAI03320.1| Hypothetical protein LOC510162 [Bo...	525	2e-147
gi|21703998|ref|NP_663482.1| aspartyl-tRNA synthetase [Mus mu...	524	3e-147
gi|28974984|ref|NP_803228.1| aspartyl-tRNA synthetase [Mus mu...	524	4e-147
gi|74181559|dbj|BAE30045.1| unnamed protein product [Mus musc...	524	4e-147
gi|74226918|dbj|BAE27102.1| unnamed protein product [Mus musculu	524	4e-147
gi|59803475|gb|AAX07827.1| cell proliferation-inducing protein 4	524	5e-147
gi|78394948|gb|AAI07750.1| Aspartyl-tRNA synthetase [Homo sapien	523	5e-147
gi|47938978|gb|AAH72534.1| Dars protein [Rattus norvegicus] >...	523	9e-147
gi|12653689|gb|AAH00629.1| Aspartyl-tRNA synthetase [Homo sap...	523	9e-147
gi|30584115|gb|AAP36306.1| Homo sapiens aspartyl-tRNA synthet...	523	9e-147
gi|55729693|emb|CAH91575.1| hypothetical protein [Pongo pygmaeus	523	1e-146

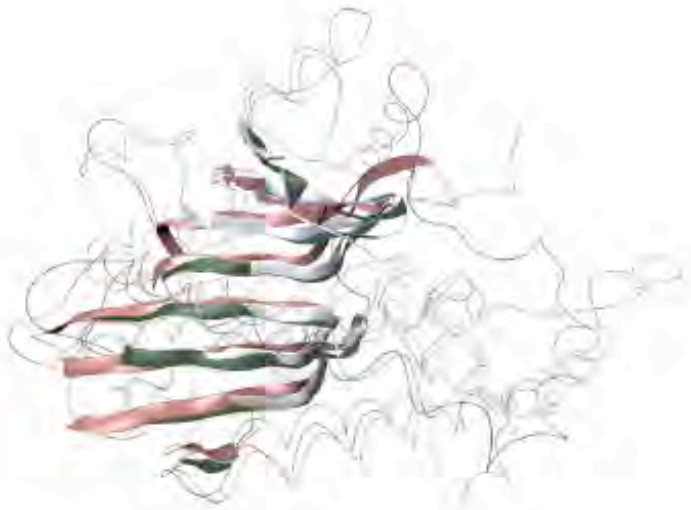
Final Result: Sequence Alignment - Approximate

```
>gi|71661457|ref|XP_817749.1|  aspartyl-tRNA synthetase [Trypanosoma cruzi strain CL Brener]
 gi|70882960|gb|EAN95898.1|  aspartyl-tRNA synthetase, putative [Trypanosoma cruzi]
Length=534

 Score =  272 bits (696),  Expect = 2e-71, Method: Composition-based stats.
 Identities = 176/507 (34%), Positives = 250/507 (49%), Gaps = 83/507 (16%)

Query  111  RARVHNTRQQGATLAFLTLRQQASLIQGLVKANKEGTISKNMVKWAGSLNLESIVLVRGI  170
            R RV  TR +G  +AF+ LRQ       +V +      + ++V+    L  ESI+   G
Sbjct  50   RGRVETTRVRG-KIAFIHLRQPPCHSIQVVAS------AADIVRRVKELTPESIIDATGT  102

Query  171  VKKVDEPIKSATVQNLEIHITKIYTISETPEALPILLEDASRSeeeeeeGLPVVNLDTR  230
            +   + P+ SA+ +N E+H  ++  +S     LP  ++D +                LDTR
Sbjct  103  LVPAERPVTSASCKNYELHAERVDVVSRAATPLPFPIKDCN-------------TRLDTR  149

Query  231  LDYRVIDLRTVTNQAIFRIQAGVCELFREYLATKKFTEVHTPKLLGAPSEGGSSVFEVTY  290
            L++RV+D+RT    ++ R+ + VC+ FR+ L  + F E+HTPK+LGA SEGGS+VF + Y
Sbjct  150  LNHRVVDMRTPLTASVLRLVSAVCQCFRKQLLARDFVEIHTPKMLGAASEGGSAVFTIDY  209

Query  291  FKGKAYLAQSPQFNKQQLIVADFERVYEIGPVFRAENSNTHRHMTEFTGLDMEMAFEEHY  350
            F  + YLAQSPQ  KQ +++ D  RV+EIGPVFRAE S THRH+TEF GLD E   E  Y
Sbjct  210  FGQRGYLAQSPQLYKQMVLMGDAMRVFEIGPVFRAEKSLTHRHLTEFVGLDAEFVIEHSY  269

Query  351  HEVLDTLSELFVFIFSELPKRFAHEIELVRKQYPVEEFKLPKDGKMVRLTYKEGIEMLRA  410
             EVLD L      +   L +  A  +   R+      +      + R   KE   ++
Sbjct  270  TEVLDVLESTVCAMIDHLQEDHAALVRQARESLADMDAAEGSPSALGRNQEKEEASIVCE  329

Query  411  AGKE-IGDFEDLSTENEK----------------------------------------FL  429
              +E +G F  L+T+  +                                        L
Sbjct  330  LSEETLGAFGCLTTDATEAHLTTDCYHGRVGVGSIDTQRRNPRQPKVLRLTFDDAVRLLL  389

Query  430  GKLVRDKYDTDF---------------YILDKFPLEIRPFYTMP-----DPANPKYSNSY  469
               V D+  TDF                Y +D + ++  P    P       P   + + S+
Sbjct  390  DHHVVDQPPTDFSLPQERRIGELVRERYGVDVYIIDQFPLTARPFYTLPHPHKTESTCSF  449

Query  470  DFFMRGEEILSGAQRIHDHALLQERMKAHGLSPEDPGLKDYCDGFSYGCPPHAGGGIGLE  529
            D ++RGEEI SGAQRIHD  LL + M+   L  +   LKDY D F YG  PH G G+GLE
Sbjct  450  DMYLRGEEICSGAQRIHDITLLLQNMER--LQVDAASLKDYVDAFRYGAWPHGGFGLGLE  507

Query  530  RVVMFYLDLKNIRRASLFPRDPKRLRP  556
            R+V+F L  K+IR+ SLFPRDPKRL P
Sbjct  508  RIVLFLLGAKDIRQISLFPRDPKRLAP  534
```

University of Illinois at Urbana-Champaign
Luthey-Schulten Group
Theoretical and Computational Biophysics Group
Summer School 2004 - University of Western Australia, Perth

Sequence Alignment Algorithms

Tutorial for the material of this lecture available

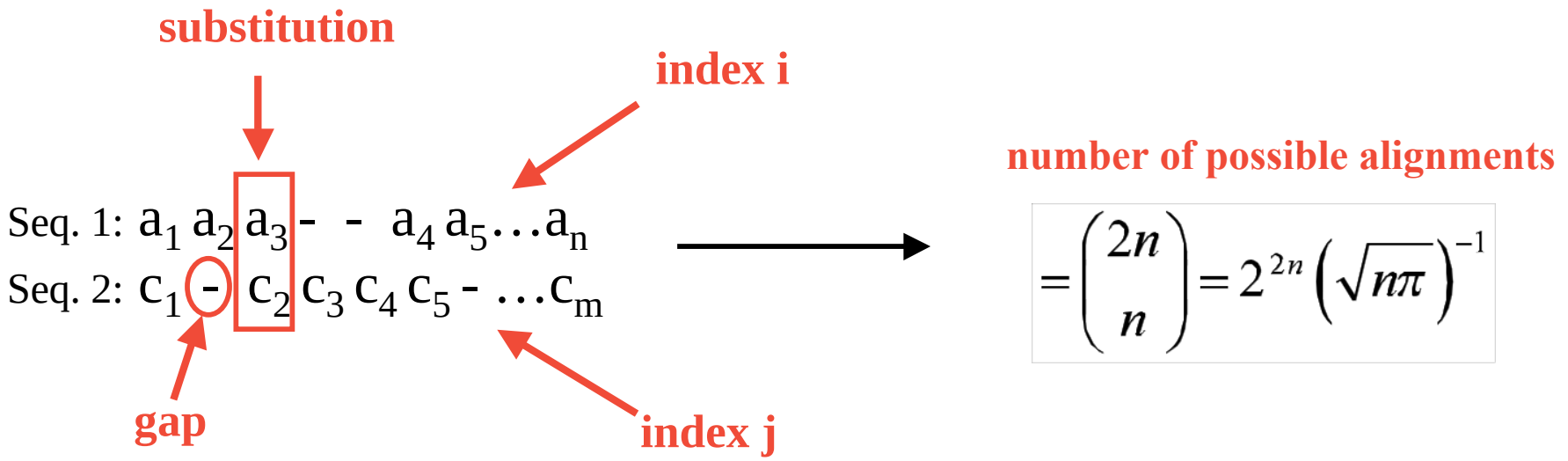


Rommie Amaro
Felix Autenrieth
Brijeet Dhaliwal
Barry Isralewitz

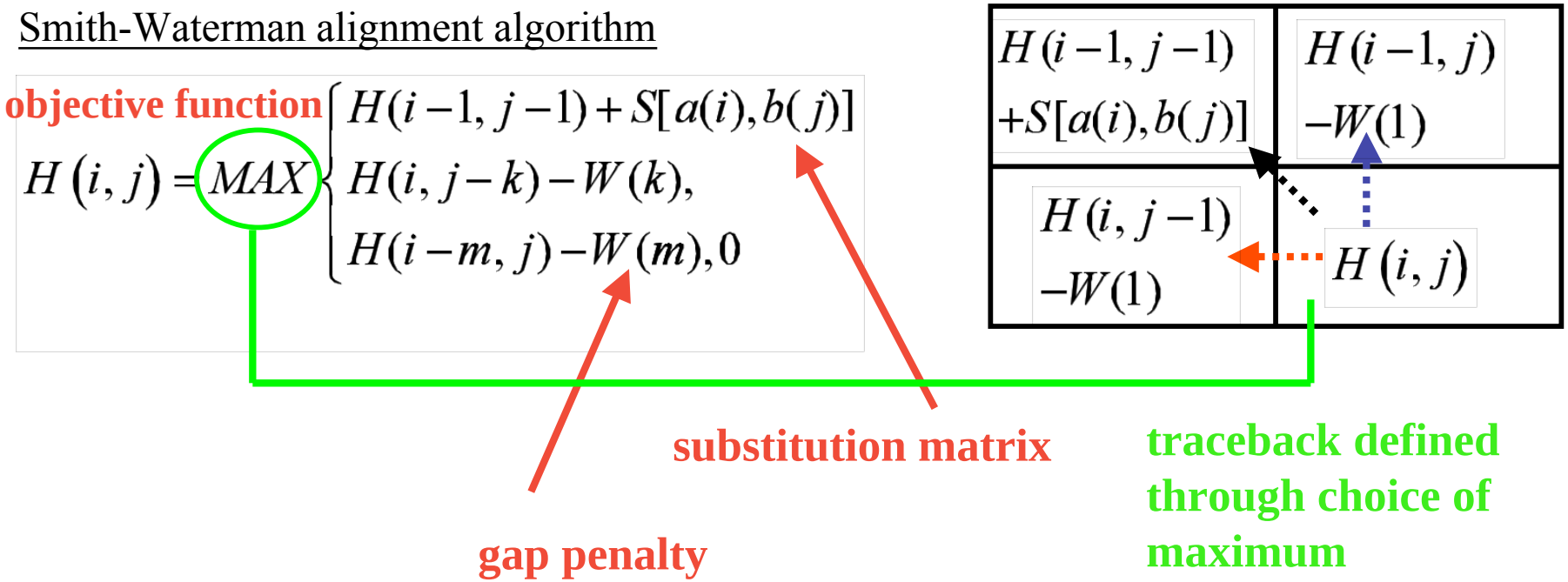
Zaida Luthey-Schulten
Anurag Sethi
Taras Pogorelov

June 2004

Sequence Alignment & Dynamic Programming



Smith-Waterman alignment algorithm



Blosum 30 Substitution Matrix

AA not resolved

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	0	0	-3	1	0	0	-2	0	-1	0	1	-2	-1	1	1	-5	-4	1	0	0	0	-7
R	-1	8	-2	-1	-2	3	-1	-2	-1	-3	-2	1	0	-1	-1	-1	-3	0	0	-1	-2	0	-1	-7
N	0	-2	8	1	-1	-1	-1	0	-1	0	-2	0	0	-1	-3	0	1	-7	-4	-2	4	-1	0	-7
D	0	-1	1	9	-3	-1	1	-1	-2	-4	-1	0	-3	-5	-1	0	-1	-4	-1	-2	5	0	-1	-7
C	-3	-2	-1	-3	17	-2	1	-4	-5	-2	0	-3	-2	-3	-3	-2	-2	-2	-6	-2	-2	0	-2	-7
Q	1	3	-1	-1	-2	8	2	-2	0	-2	-2	0	-1	-3	0	-1	0	-1	-1	-3	-1	4	0	-7
E	0	-1	-1	1	1	2	6	-2	0	-3	-1	2	-1	-4	1	0	-2	-1	-2	-3	0	5	-1	-7
G	0	-2	0	-1	-4	-2	-2	8	-3	-1	-2	-1	-2	-3	-1	0	-2	1	-3	-3	0	-2	-1	-7
H	-2	-1	-1	-2	-5	0	0	-3	14	-2	-1	-2	2	-3	1	-1	-2	-5	0	-3	-2	0	-1	-7
I	0	-3	0	-4	-2	-2	-3	-1	-2	6	2	-2	1	0	-3	-1	0	-3	-1	4	-2	-3	0	-7
L	-1	-2	-2	-1	0	-2	-1	-2	-1	2	4	-2	2	2	-3	-2	0	-2	3	1	-1	-1	0	-7
K	0	1	0	0	-3	0	2	-1	-2	-2	-2	4	2	-1	1	0	-1	-2	-1	-2	0	1	0	-7
M	1	0	0	-3	-2	-1	-1	-2	2	1	2	2	6	-2	-4	-2	0	-3	-1	0	-2	-1	0	-7
F	-2	-1	-1	-5	-3	-3	-4	-3	-3	0	2	-1	-2	10	-4	-1	-2	1	3	1	-3	-4	-1	-7
P	-1	-1	-3	-1	-3	0	1	-1	1	-3	-3	1	-4	-4	11	-1	0	-3	-2	-4	-2	0	-1	-7
S	1	-1	0	0	-2	-1	0	0	-1	-1	-2	0	-2	-1	-1	4	2	-3	-2	-1	0	-1	0	-7
T	1	-3	1	-1	-2	0	-2	-2	-2	0	0	-1	0	-2	0	2	5	-5	-1	1	0	-1	0	-7
W	-5	0	-7	-4	-2	-1	-1	1	-5	-3	-2	-2	-3	1	-3	-3	-5	20	5	-3	-5	-1	-2	-7
Y	-4	0	-4	-1	-6	-1	-2	-3	0	-1	3	-1	-1	3	-2	-2	-1	5	9	1	-3	-2	-1	-7
V	1	-1	-2	-2	-2	-3	-3	-3	-3	4	1	-2	0	1	-4	-1	1	-3	1	5	-2	-3	0	-7
B	0	-2	4	5	-2	-1	0	0	-2	-2	-1	0	-2	-3	-2	0	0	-5	-3	-2	5	0	-1	-7
Z	0	0	-1	0	0	4	5	-2	0	-3	-1	1	-1	-4	0	-1	-1	-1	-2	-3	0	4	0	-7
X	0	-1	0	-1	-2	0	-1	-1	-1	0	0	0	0	-1	-1	0	0	-2	-1	0	-1	0	-1	-7
*	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	1

Amino Acid Three Letter and One Letter Code

Amino acid	Three letter code	One letter code
alanine	ala	A
arginine	arg	R
asparagine	asn	N
aspartic acid	asp	D
asparagine or aspartic acid	asx	B
cysteine	cys	C
glutamic acid	glu	E
glutamine	gln	Q
glutamine or glutamic acid	glx	Z
glycine	gly	G
histidine	his	H
isoleucine	ile	I
leucine	leu	L
lysine	lys	K
methionine	met	M
phenylalanine	phe	F
proline	pro	P
serine	ser	S
threonine	thr	T
tryptophan	try	W
tyrosine	tyr	Y
valine	val	V

Sequence Alignment & Dynamic Programming

Seq. 1: $a_1\ a_2\ a_3$ - - $a_4\ a_5 \ldots a_n$

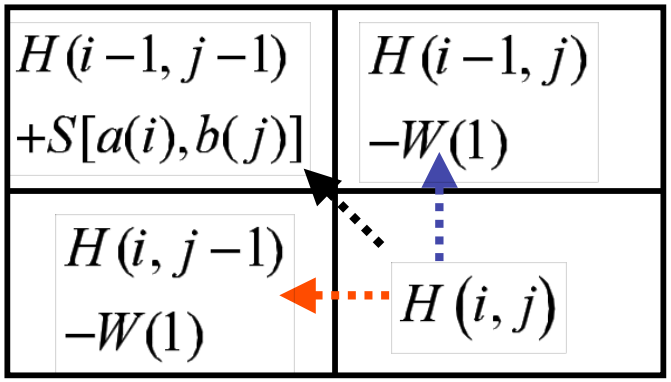
Seq. 2: c_1 - $c_2\ c_3\ c_4\ c_5$ - $\ldots c_m$

number of possible alignments:

$$= \binom{2n}{n} = 2^{2n}\left(\sqrt{n\pi}\right)^{-1}$$

Needleman-Wunsch alignment algorithm

$$H(i,j) = MAX\begin{cases} H(i-1,j-1) + S[a(i),b(j)] \\ H(i,j-k) - W(k), \\ H(i-m,j) - W(m) \end{cases}$$



Score Matrix H: Traceback

S : substitution matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V		B	Z	X	
A	5	-2	-1	-1	-2	0	-1	1	-2	-1	-2	-1	-1	-3	-2	1	0	-3	-2	0	-1	-1	0	A
R	-2	9	0	-1	-3	2	-1	-3	0	-3	-2	3	-1	-2	-3	-1	-2	-2	-1	-2	-1	0	-1	R
N	-1	0	8	2	-2	1	-1	0	1	-2	-3	0	-2	-3	-2	1	0	-4	-2	-3	4	0	-1	N
D	-1	-1	2	9	-2	-1	2	-2	0	-4	-3	0	-3	-4	-2	0	-1	-5	-3	-3	6	1	-1	D
C	-2	-3	-2	-2	16	-4	-2	-3	-4	-4	-2	-3	-3	-2	-5	-1	-1	-6	-4	-2	-2	-3	-2	C
Q	0	2	1	-1	-4	8	2	-2	0	-3	-2	1	-1	-4	-2	1	-1	-1	-1	-3	0	4	-1	Q
E	-1	-1	-1	2	-2	2	7	-3	0	-4	-2	1	-2	-3	0	0	-1	-2	-2	-3	1	5	-1	E
G	1	-3	0	-2	-3	-2	-3	8	-2	-4	-4	-2	-2	-3	-1	0	-2	-2	-3	-4	-1	-2	-1	G
H	-2	0	1	0	-4	0	0	-2	13	-3	-2	-1	1	-2	-2	-1	-2	-5	2	-4	0	0	-1	H
I	-1	-3	-2	-4	-4	-3	-4	-4	-3	6	2	-3	1	1	-2	-2	-1	-3	0	4	-3	-4	-1	I
L	-2	-2	-3	-3	-2	-2	-2	-4	-2	2	6	-2	3	2	-4	-3	-1	-1	0	2	-3	-2	-1	L
K	-1	3	0	0	-3	1	1	-2	-1	-3	-2	6	-1	-3	-1	0	0	-2	-1	-2	0	1	-1	K
M	-1	-1	-2	-3	-3	-1	-2	-2	1	1	3	-1	7	0	-2	-2	-1	-2	1	1	-3	-2	0	M
F	-3	-2	-3	-4	-2	-4	-3	-3	-2	1	2	-3	0	9	-4	-2	-1	1	4	0	-3	-4	-1	F
P	-2	-3	-2	-2	-5	-2	0	-1	-2	-2	-4	-1	-2	-4	11	-1	0	-4	-3	-3	-2	-1	-2	P
S	1	-1	1	0	-1	1	0	0	-1	-2	-3	0	-2	-2	-1	5	2	-5	-2	-1	0	0	0	S
T	0	-2	0	-1	-1	-1	-1	-2	-2	-1	-1	0	-1	-1	0	2	6	-4	-1	1	0	-1	0	T
W	-3	-2	-4	-5	-6	-1	-2	-2	-5	-3	-1	-2	-2	1	-4	-5	-4	19	3	-3	-4	-2	-2	W
Y	-2	-1	-2	-3	-4	-1	-2	-3	2	0	0	-1	1	4	-3	-2	-1	3	9	-1	-3	-2	-1	Y
V	0	-2	-3	-3	-2	-3	-3	-4	-4	4	2	-2	1	0	-3	-1	1	-3	-1	5	-3	-3	-1	V
B	-1	-1	4	6	-2	0	1	-1	0	-3	-3	0	-3	-3	-2	0	0	-4	-3	-3	5	2	-1	B
Z	-1	0	0	1	-3	4	5	-2	0	-4	-2	1	-2	-4	-1	0	-1	-2	-2	-3	2	5	-1	Z
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	0	-1	-2	0	0	-2	-1	-1	-1	-1	-1	X

gap penalty W = - 6

Needleman-Wunsch Global Alignment

Similarity Values

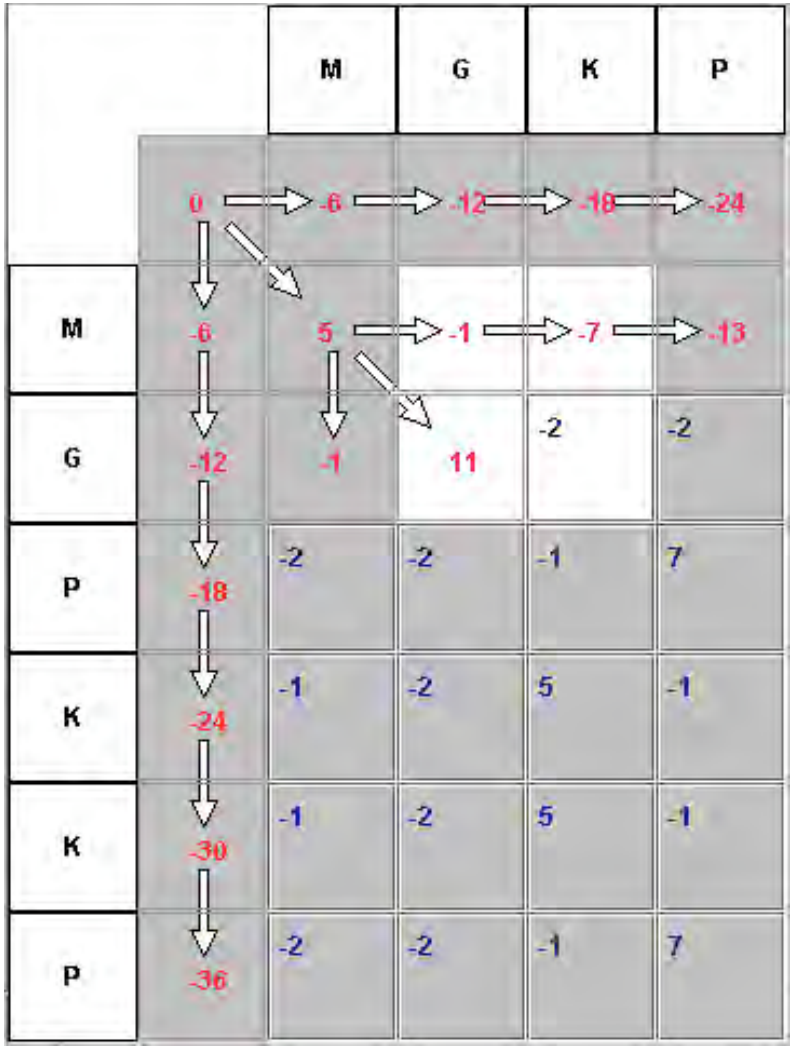
		M	G	K	P
M		5	-3	-1	-2
G		-3	6	-2	-2
P		-2	-2	-1	7
K		-1	-2	5	-1
K		-1	-2	5	-1
P		-2	-2	-1	7

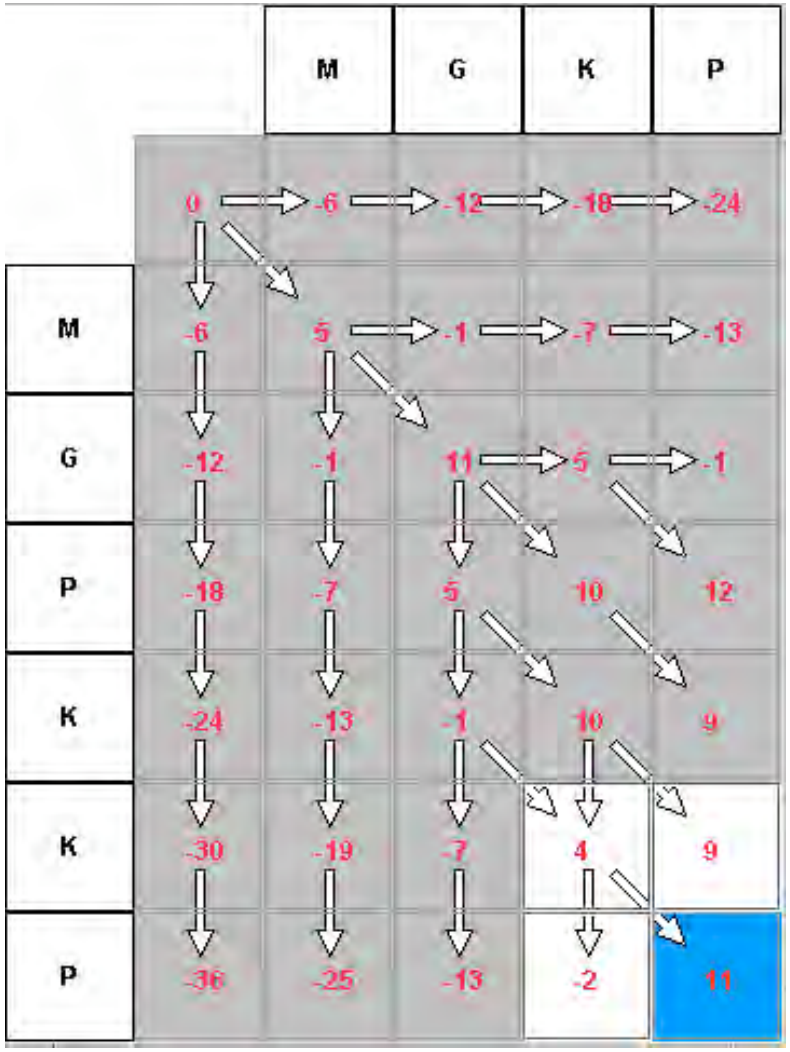
Initialization of Gap Penalties

		M	G	K	P
	0	-6	-12	-18	-24
M	-6	5	-3	-1	-2
G	-12	-3	6	-2	-2
P	-18	-2	-2	-1	7
K	-24	-1	-2	5	-1
K	-30	-1	-2	5	-1
P	-36	-2	-2	-1	7

http://www.dkfz-heidelberg.de/tbi/bioinfo/PracticalSection/AliApplet/index.html

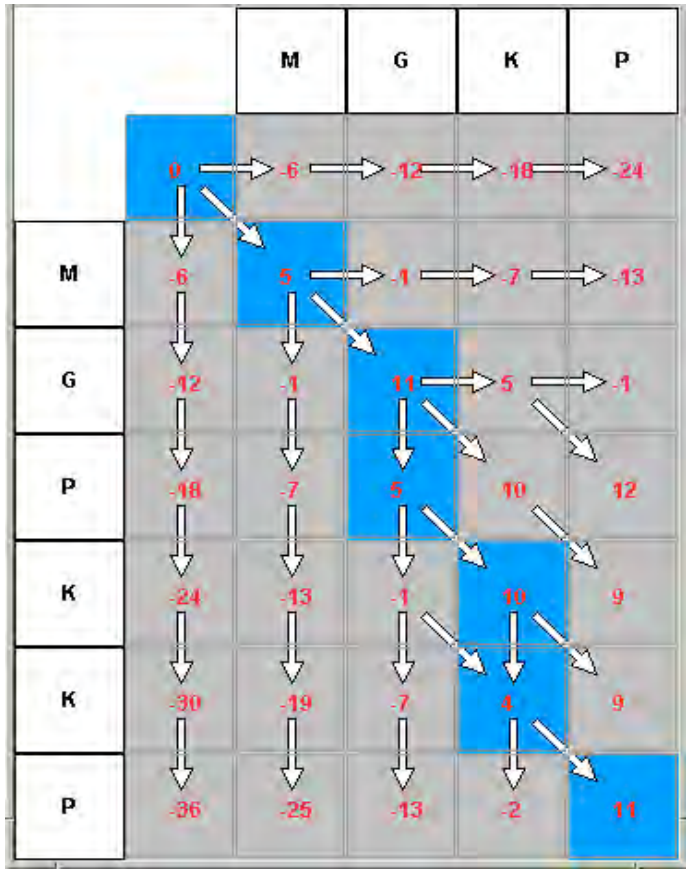
Filling out the Score Matrix H





http://www.dkfz-heidelberg.de/tbi/bioinfo/PracticalSection/AliApplet/index.html

Traceback and Alignment



The Alignment

M	G	-	K	-	P
:	:		:		:
M	G	P	K	K	P

Traceback (blue) from optimal score

http://www.dkfz-heidelberg.de/tbi/bioinfo/PracticalSection/AliApplet/index.html

Protein Structure Prediction

1-D protein sequence

SISSIRVKSKRIQLG…

3-D protein structure

Homology Modeling/ FR

$$E = E_{match} + E_{gap}$$

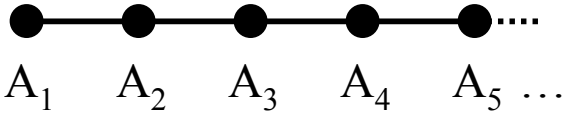
Target Sequence

SISSRVKSKRIQLGLNQAELAQKV------GTTQ…
QFANEFKVRRIKLGYTQ----TNVGEALAAVHGS…

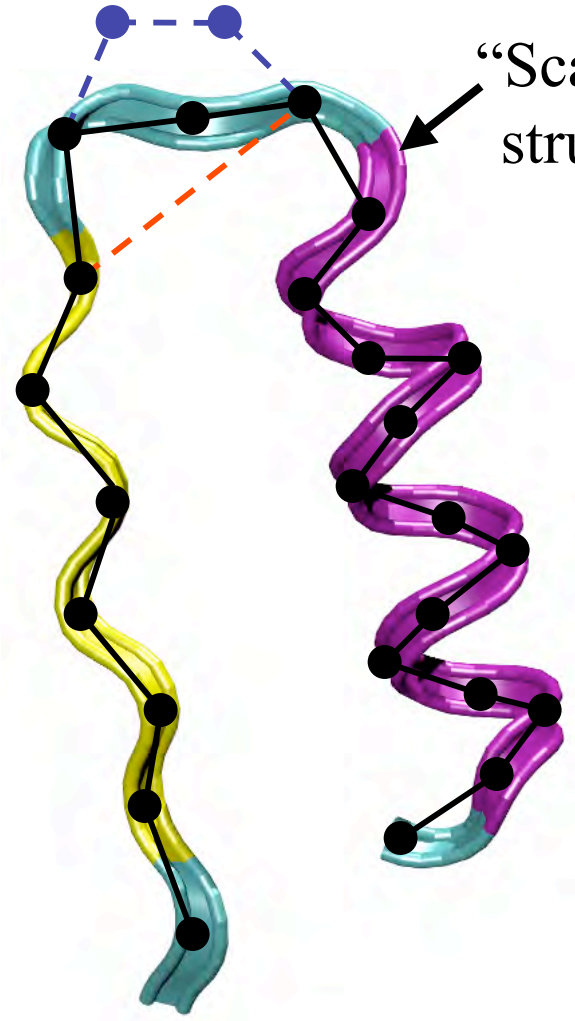
Known structure(s)

Sequence-Structure Alignment

Target sequence



Alignment between target(s) and scaffold(s)



1. Energy Based Threading*

$$H = E_{contact} + E_{profile} + E_{H-bonds} + E_{gap}$$

$$E_{profile} = \sum_{i}^{n} \gamma^{(p)}\left(A_i, SS_i, SA_i\right)$$

$$E_{contact} = \sum_{i,j} \sum_{k=1}^{2} \gamma_k^{(ct)}\left(A_i, A_j\right) * U\left(r_k - r_{ij}\right)$$

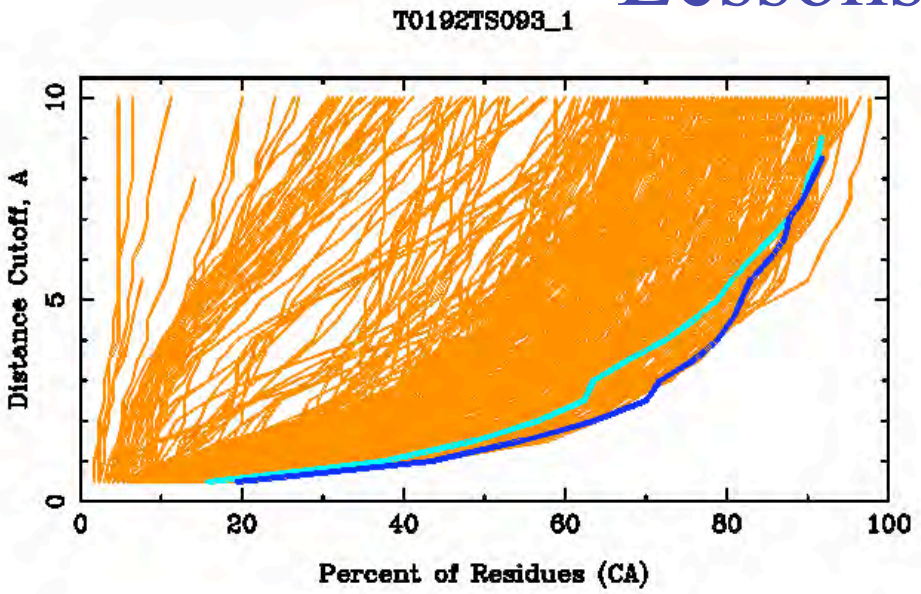
2. Sequence – Structure Profile Alignments

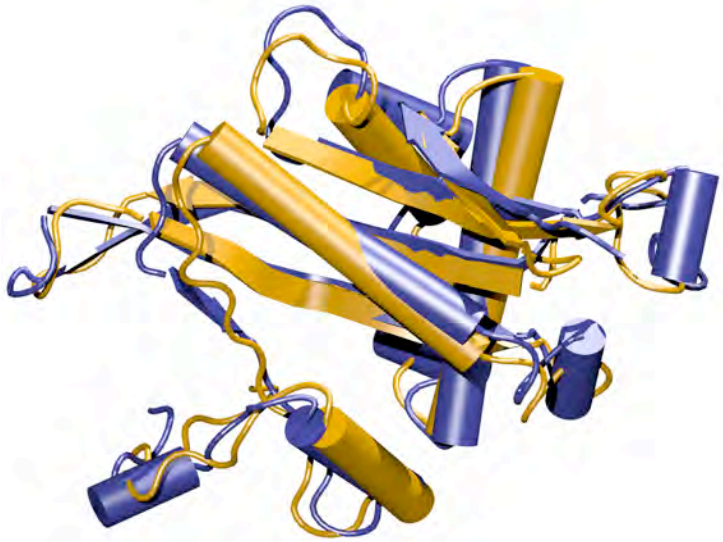
Clustal, Hidden Markov (HMMER, PSSM)
with position dependent gap penalties

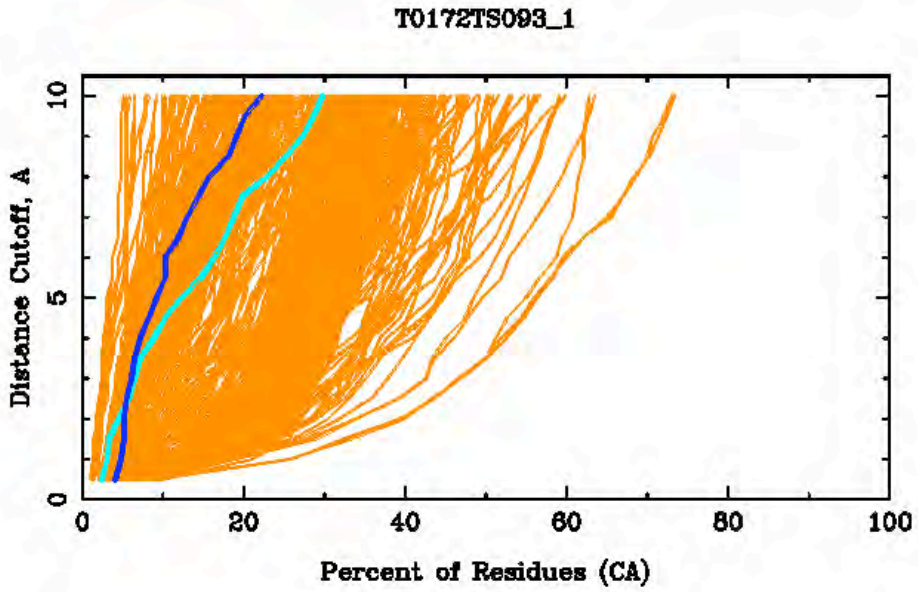
*R. Goldstein, Z. Luthey-Schulten, P. Wolynes (1992, PNAS), K. Koretke et.al. (1996, Proteins)

CM/Fold Recognition Results from CASP5

Lessons Learned







The prediction is never better than the scaffold.

Threading Energy Function and Profiles need improvement.

Structural Profiles

1. Structure more conserved than sequences!!! Similar structures at the Family and Superfamily levels.

   Add more structural information

2. Which structures and sequences to include? Use evolution and eliminate redundancy with QR factorization

Structural Domains

Structural Classification of Proteins

Protein: Aspartyl-tRNA synthetase (AspRS) from *Escherichia coli*

Lineage:

1. Root: scop
2. Class: All beta proteins
3. Fold: OB-fold
   barrel, closed or partly opened n=5, S=10 or S=8; greek-key
4. Superfamily: Nucleic acid-binding proteins
5. Family: Anticodon-binding domain
   barrel, closed; n=5, S=10
6. Protein: Aspartyl-tRNA synthetase (AspRS)
   this is N-terminal domain in prokaryotic enzymes and the first "visible" domain in eukaryotic enzymes
7. Species: *Escherichia coli*

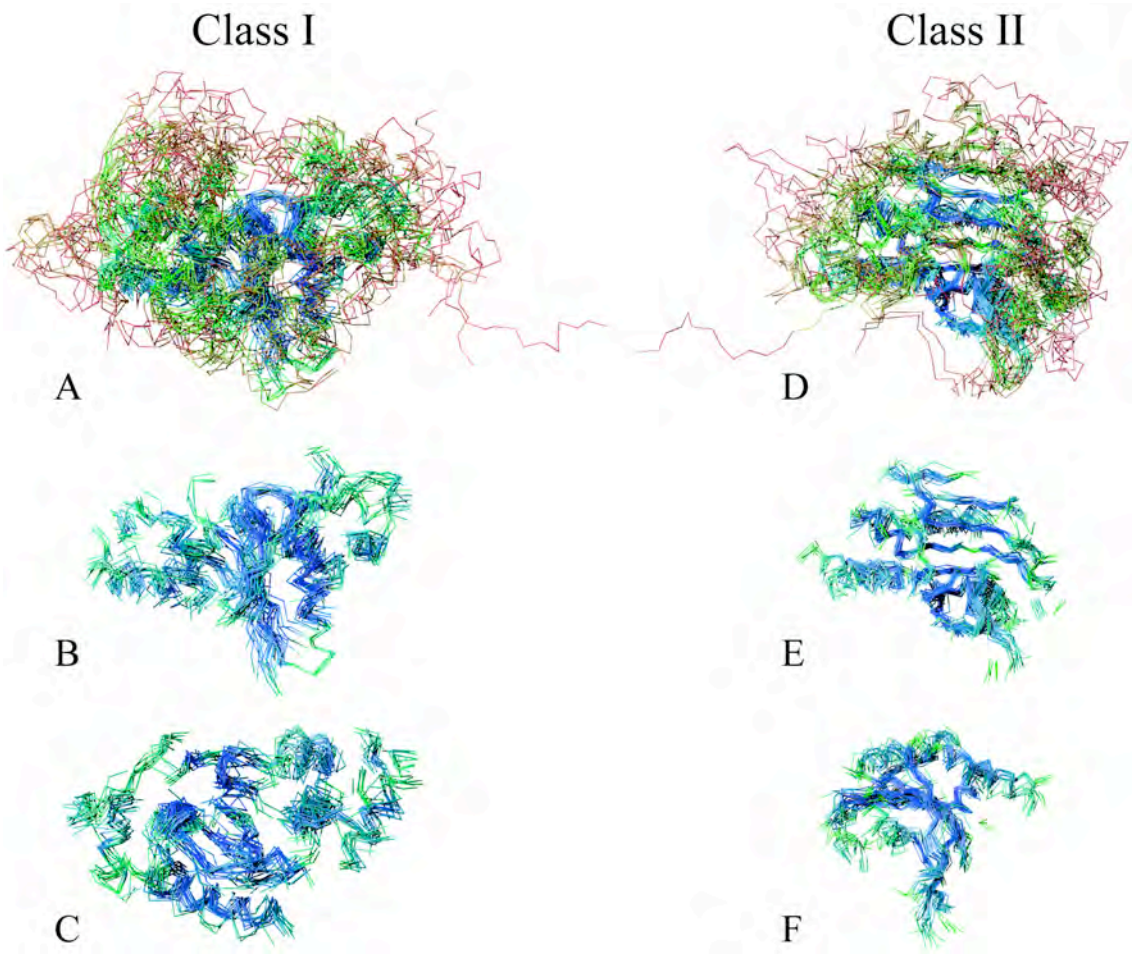
PDB Entry Domains:

1. 1c0a
   1. region a:1-106
2. 1il2
   complexed with 1mg, 5mc, 5mu, amo, h2u, psu, so4
   1. region a:1-106
   2. region b:1001-1106
3. 1eqr
   complexed with mg
   1. region a:1-106
   2. region b:1-106
   3. region c:1-106

Profile - Multiple Structural Alignments

Representative Profile of AARS Family

Catalytic Domain



STAMP - Multiple Structural Alignments

1. Initial Alignment Inputs

- Multiple Sequence alignment
- Ridged Body "Scan"

2. Refine Initial Alignment & Produce Multiple Structural Alignment

$$P_{ij} = \left\{ e^{-d_{ij}^2 / 2E_1} \right\} \left\{ e^{-s_{ij}^2 / 2E_2} \right\}$$

probability that residue i on structure A is equivalent to residue j on structure B.

d_{ij} -- distance between i & j

S_{ij} -- conformational similarity; function of rms bewteen i-1, i, i+1 and j-1, j, j+1.

- Dynamic Programming (Smith-Waterman) through P matrix gives optimal set of equivalent residues.
- This set is used to re-superpose the two chains. Then iterate until alignment score is unchanged.
- This procedure is performed for all pairs.

R. Russell, G. Barton (1992) *Proteins* **14:** 309.

Multiple Structural Alignments

STAMP – cont'd

2. Refine Initial Alignment & Produce Multiple Structural Alignment

Alignment score:

$$S_C = \frac{S_p}{L_P}\frac{L_p - i_A}{L_A}\frac{L_p - i_B}{L_B}$$

$$S_p = \sum_{aln.path} P_{ij}$$

L_p, L_A, L_B -- length of alignment, sequence A, sequence B

i_A, i_B -- length of gaps in A and B.

Multiple Alignment:

- Create a dendrogram using the alignment score.
- Successively align groups of proteins (from branch tips to root).
- When 2 or more sequences are in a group, then average coordinates are used.

Variation in Secondary Structure STAMP Output

Eargle, O'Donoghue, Luthey-Schulten, UIUC 2004

R. Russell, G. Barton (1992) *Proteins* **14:** 309

Stamp Output/Clustal Format

```
SerRS-T_thermophilus          VGGEEANREIKRVGGPPEFSFP--P--LDHVALMEKNGWWEPRISQVSGSRSYALKGDLA
ThrRS-E_coli                  -------------------------R--DHRKIGKQLDLY-HMQ-EE-APGMVFWHNDGW
ProRS-T_thermophilus          ----------------KGLTPQSQDFSEWYLEVIQKAELAD-YG--P-VRGTIVVRPYGY
ProRS-M_thermoautotrophicus   ------------------------EFSEWFHNILEEAEIIDQRY--P-VKGMHVWMPHGF
space                         ------------------------------------------------------------
SerRS-T_thermophilus          --SGGG-EEEEEES-----SS--------HHHHHHHHT-B-TTHHHHH-SS---B-THHH
ThrRS-E_coli                  ----------------------------HHHHHHHTT-E-E---TT-STT--EE-HHHH
ProRS-T_thermophilus          --------------------HHHHHHHHHHHHHHHHTTSEE-E---S-STT-EEE-HHHH
ProRS-M_thermoautotrophicus   ------------------------HHHHHHHHHHHTT-EE-----S-STT--EE-HHHH

SerRS-T_thermophilus          LYELALLRFAMDFMARRGFLPMTLPSYAREK-AFLG-TGHFPAYRDQVWAIA-----E--
ThrRS-E_coli                  TIFRELEVFVRSKLKEYQYQEVKGPFMMDRV-LWEKT-GHWDNYKDAMFTTS----S-EN
ProRS-T_thermophilus          AIWENIQQVLDRMFKETGHQNAYFPLFIPMSFL----------FSPELAVVTHAGGEELE
ProRS-M_thermoautotrophicus   MIRKNTLKILRRILD-RDHEEVLFPLLVPEDE-LAKEAIHVKGFEDEVYWVTHGGLSKLQ
space                         ------------------------------------------------------------
SerRS-T_thermophilus          HHHHHHHHHHHHHHHHTT-EEEE--SEEEHH-HHHH-HT-TTTTGGGS-B-T-----T--
ThrRS-E_coli                  HHHHHHHHHHHHHHHHHTT-EE----SEEEHH-HHHTT-THHHHHGGG--EEE----E-TT
ProRS-T_thermophilus          HHHHHHHHHHHHHHHHHTT-EE----SEEESTT-------------TT--EEEE-SSSEEE
ProRS-M_thermoautotrophicus   HHHHHHHHHHHHHHT-TT-EE----SEEEHHH-HTTSHHHHHHTTTT--EEEEETTEEEE

SerRS-T_thermophilus          TDLYLTGTAEVVLNALHSGEILPYEALPLRYAGYAPAFRSEA--GSFGKDVRGLMRVH-Q
ThrRS-E_coli                  REYCIKPMNCPGHVQIFNQGLKSYRDLPLRMAEFGSCHR--NEPS--G-SLHGLMRVR-G
ProRS-T_thermophilus          EPLAVRPTSETVIGYMWSKWIRSWRDLPQLLNQWGNVVRW--E----M-RTRPFLRTSE-
ProRS-M_thermoautotrophicus   RKLALRPTSETVMYPMFALWVRSHTDLPMRFYQVVNTFRY-ET----K-HTRPLIRVREI
space                         ------------------------------------------------------------
SerRS-T_thermophilus          SSEEE-S-THHHHHHHTTT-EEEGGG-SEEEEEEEEE------S--SSTTTTTTS-S-E
ThrRS-E_coli                  EEEEE-S-SHHHHHHHHTSS--BTTT-SEEEEE--EEE------G--G-G-BTTTB-S-E
ProRS-T_thermophilus          EEEEE-S-SHHHHHHHHHHH--BGGG--EEEEEEEEEE---------S-S-BTTTB-SE-
ProRS-M_thermoautotrophicus   EEEEE-SSSHHHHHHHHHHH--BTTT--EEEEEEEEEE---------S---BTTTB-SEE
```

From multiple structure alignment compute position probabilities for amino acids and gaps!!!!

PSSM-based approach

I. Construction of Profile

	1	2	3	4	5
Sequence 1	-	B	B	-	C
Sequence 2	C	C	-	-	C
Sequence 3	C	B	C	C	B

Multiple Sequence Alignment

↓

Position Specific Amino Acid Probabilities

j	1	2	3	4	5
P(C^j)	1	0.33	0.5	1	0.67
P(B^j)	0	0.67	0.5	0	0.33

Position specific score for aligning i^{th} residue of S to j^{th} position of profile

$$Sc(S_i^j) = \log(\mathrm{P}(S_i^j)/\mathrm{P}(S_i^{rnd}))$$

II Database search

Align every sequence in the database to the profile using Dynamic Programming algorithm.

Sequence represented by $S(S_1,S_2,\ldots,S_{n_{res}})$

Progressive alignment score = *H(i,j)*

$$H(0,0) = 0, H(i,0) = i \times \delta, H(0,j) = j \times \delta.$$

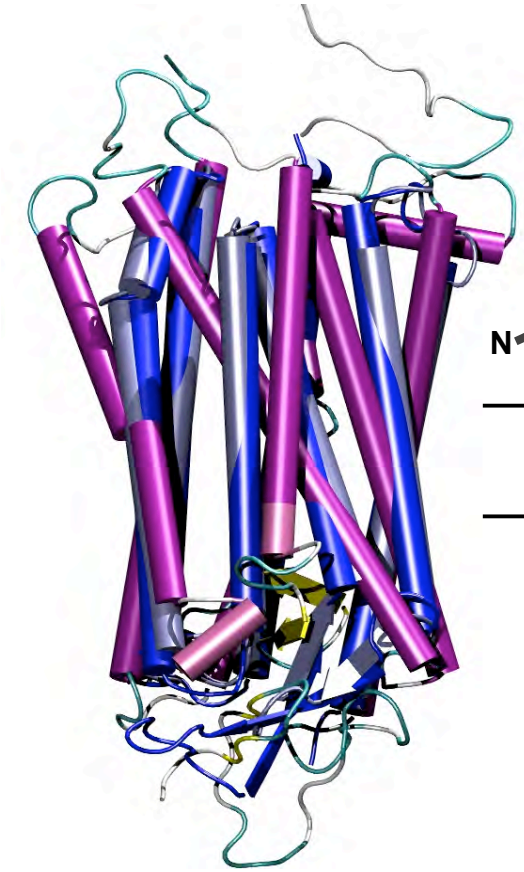
$$H(i,j) = max \begin{cases} H(i-1,j-1) + Sc(S_i^j) \\ H(i-1,j) + \delta \\ H(i,j-1) + \delta \end{cases}$$

for $j=1,2,\ldots,m_{aln}$ and $i=1,2,\ldots,N_{res}$

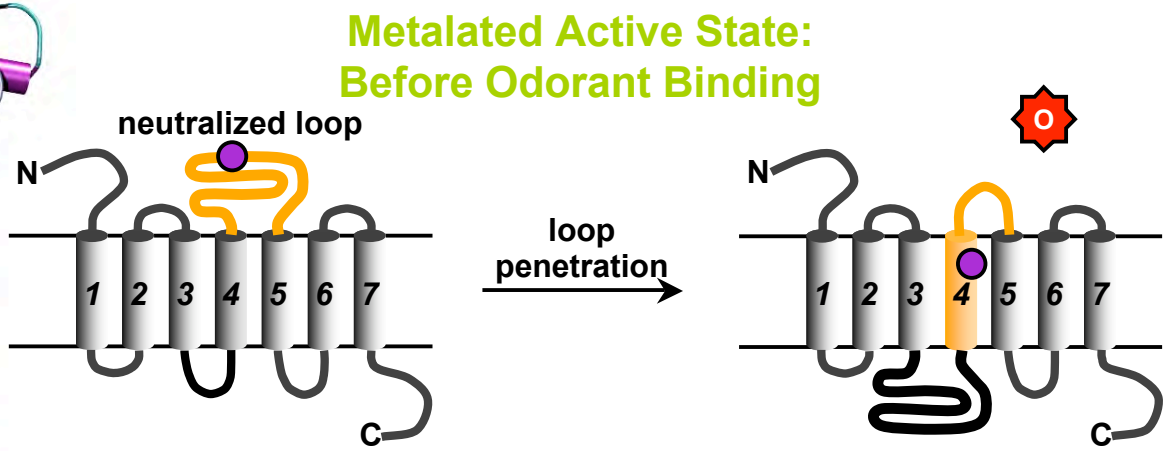
Traceback gives the optimal alignment of the sequence S to profile.

Altschul, et. al., *Nucleic Acids Research*, **25:**3389-3402. Also useful: http://apps.bioneq.qc.ca/twiki/bin/view/Knowledgebase/PSSM

HMM / Clustal Models of Transmembrane Proteins



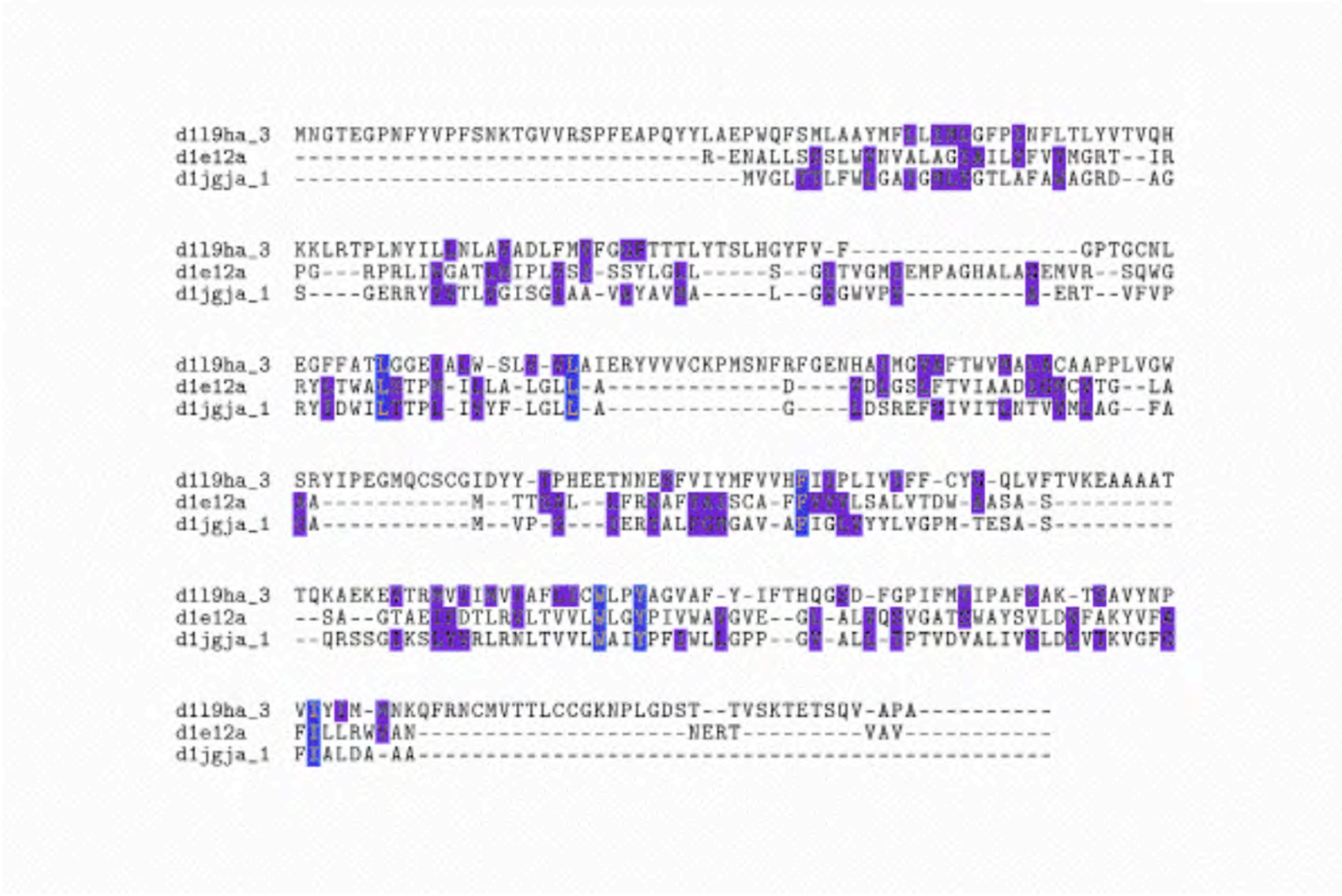
Bacteriorhodpsin/Rhodopsins



Olfactory Receptor/Bovine Rhodopsin

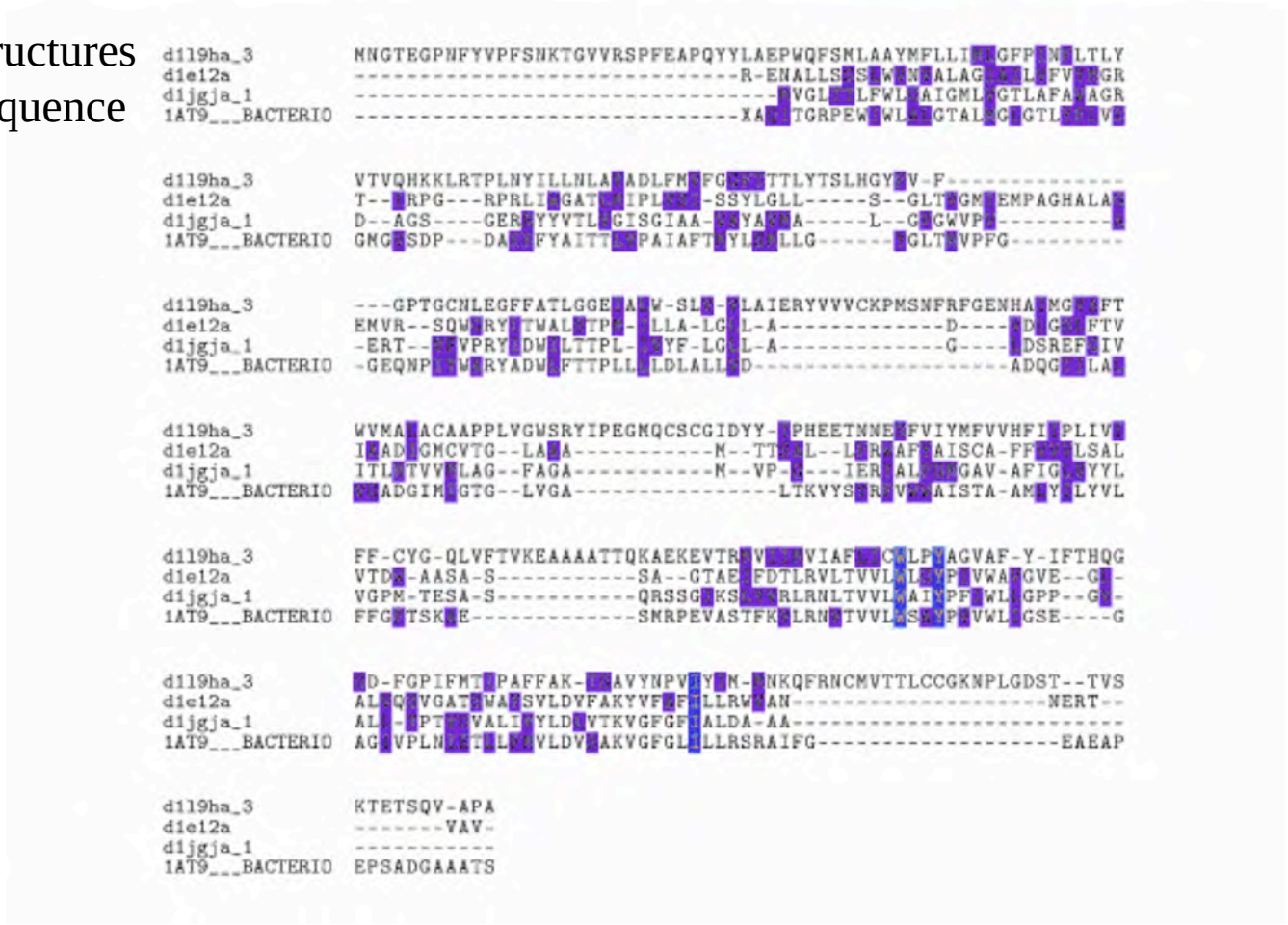
J. Wang, Z. Luthey-Schulten, K. Suslick (2003) *PNAS* **100**(6):3035-9

Stamp Profile - 3 Structures



Clustal Profile-Profile Alignment

Profile 1 Structures
Profile 2 Sequence



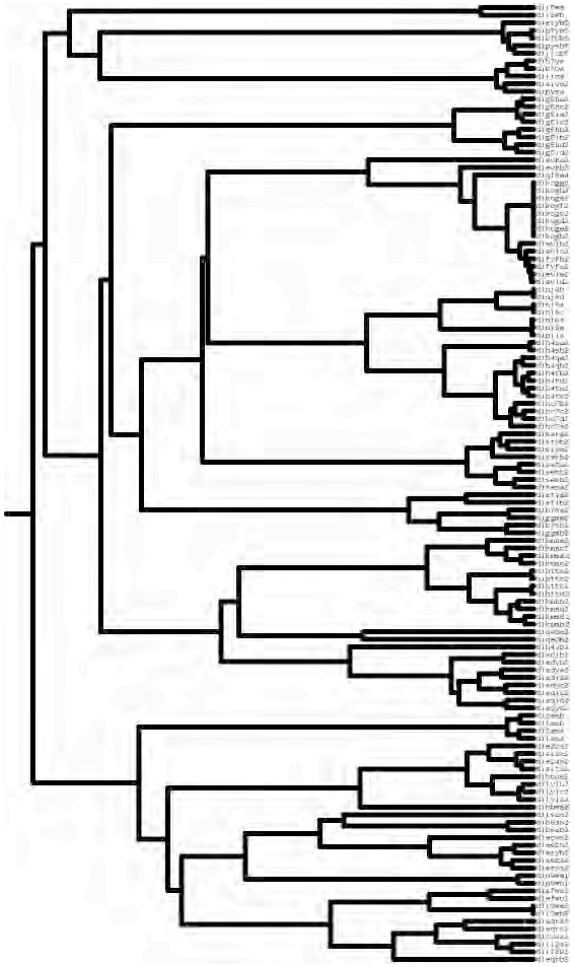
Refine Structure Prediction with Modeller 6.2

Sethi and Luthey-Schulten, UIUC 2003

Modeller 6.2 A. Sali, et al.

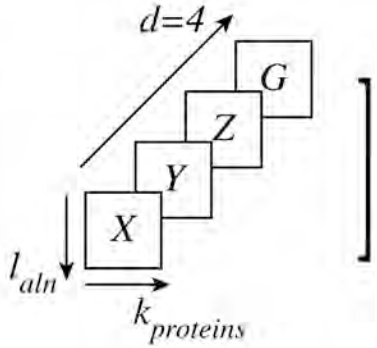
Non-redundant Representative Sets

Too much information
129 Structures

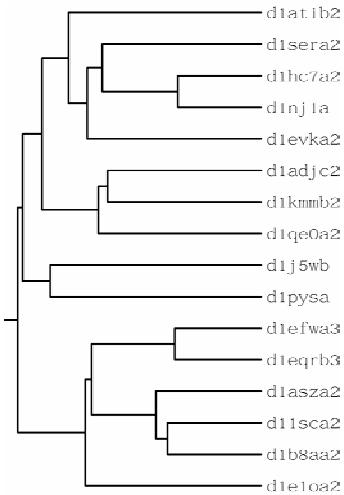


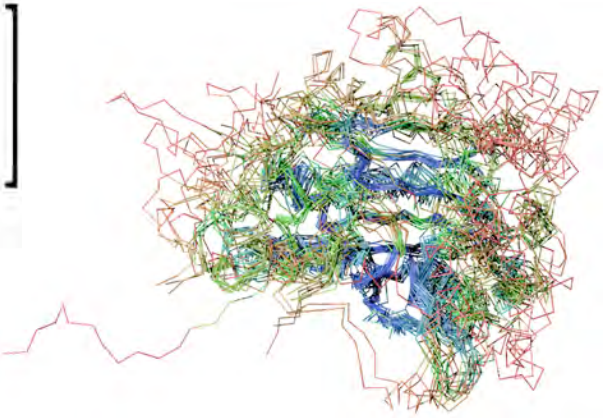
Multidimensional QR factorization of alignment matrix, *A*.

$$A = \left[\begin{array}{c} \\ \\ \\ \\ \end{array} \right]$$



Economy of information
16 representatives





QR computes a set of maximal linearly independent structures.

P. O'Donoghue and Z. Luthey-Schulten (2003) *MMBR* **67**:550-571.

P. O'Donoghue and Z. Luthey-Schulten (2005) *J. Mol. Biol.*, **346**, 875-894.

On to Evolution of Protein Structure