



# Mining the Human Genome Using Protein Structure Homology

Randal R. Ketchum, Ph.D.  
Amgen Inc.

**AMGEN**

WASHINGTON

# Introduction

Need for gene mining

Scale of problem

Protein structure

Structure prediction

Mining the genome

Some results

Some problems

Future work

```
TCTCGAGGGCCACGCGTTTAAACGTTCGAGGTACCTATCCCGGGCCGCCAC
CATGGCTACAGGCTCCCGGACGTCCCTGCTCCTGGCTTTTGGCCTGCTCT
GCCTGCCCTGGCTTCAAGAGGGCAGTGCAACTAGTTCTGACCGTATGAAA
CAGATAGAGGATAAGATCGAAGAGATCCTAAGTAAGATTTATCATATAGA
GAATGAAATCGCCCGTATCAAAAAGCTGATTGGCGAGCGGACTAGATCTA
GTTTGGGGAGCCGGGCATCGCTGTCCGCCAGGAGCCTGCCCAGGAGGAG
CTGGTGGCAGAGGAGGACCAGGACCCGTCGGAACCTGAATCCCAGACAGA
AGAAAGCCAGGATCCTGCGCCTTTCCTGAACCGACTAGTTCCGGCCTCGCA
AAGTGCACCTAAAGGCCGAAAACACGGGCTCGAAGAGCGATCGCAGCC
CATTATGAAGTTCATCCACGACCTGGACAGGACGGAGCGCAGGCAGGTGT
GGACGGGACAGTGAGTGGCTGGGAGGAAGCCAGAATCAACAGCTCCAGCC
TCTGCGCTACAACCGCCAGATCGGGGAGTTTATAGTCACCCGGGCTGGG
CTCTACTACCTGTACTGTCAGGTGCACTTTGATGAGGGGAAGGCTGTCTA
CCTGAAGCTGGACTTGCTGGTGGATGGTGTGCTGGCCCTGCGCTGCCTGG
AGGAATTCTCAGCCACTGCGGCCAGTTCCTCGGGCCCCAGCTCCGCCTC
TGCCAGGTGTCTGGGCTGTTGGCCCTGCGGCCAGGGTCCCTCCCTGCGGAT
CCGCACCCTCCCCTGGGCCATCTCAAGGCTGCCCCCTTCCTCACCTACT
TCGGACTCTTCCAGGTTCACTGAGCGGCCGCGGATCTGTTTAAACTAG
```

```
MATGSRTSLLLAFGLLCLPWLQEGSATSSDRMKQIEDKIEEILSKIYHIE
NEIARIKKLIGERTRSSLGSRASLSAQEPAQEELVAEEDQDPSELNPQTE
ESQDPAPFLNRLVRRRSAPKGRKTRARRAIAAHYEVHPRPGDGAQAGV
DGTVSGWEEARINSSSPLRYNRQIGEFIVTRAGLYYLYCQVHFDEGKAVY
LKLDDLVDGVLALRCLLEEFSAATAASSLGPQLRLCQVSGLLALRPGSSLRI
RTLPAHLKAAPFLTYFGLFQVH
```

**AMGEN**

WASHINGTON

## **Need For Gene Mining**

**Human Genome contains approximately  
30-60 thousand genes**

**Only 30-40% of these are classified into  
known function families**

**Function of proteins needed to enable  
development of therapeutics**

**AMGEN**

WASHINGTON

# Need For Gene Mining

Experimental methods too slow for complete classification

Computational methods for elucidating function needed

Weeks or months, around \$100K, to experimentally solve single, globular structure

**AMGEN**

WASHINGTON

# **NIGMS Structural Genomics Initiative**

**Proteins fold into a limited number of shapes**

**Estimates of ~10K protein folds - ~700 currently in the PDB**

**Solve key structures within families - homology can be used for rest**

**Around 10 years to solve 10K unique structures**

**Problem - many proteins have same fold with little or no sequence homology**

**AMGEN**

WASHINGTON

## Scale of the Problem

~15K structures in the Protein Data Bank

Around 4K are unique (< 90% identical)

This represents ~1500 families and ~700 folds

Less than 10% of all chains discovered in 2001 were new folds

So - many genes are for unknown function with no hope of change in the near future

**AMGEN**

WASHINGTON

# SCOP Family

Family: Short-chain cytokines

Lineage:

1. Root: scop
2. Class: All alpha proteins
3. Fold: 4-helical cytokines  
core: 4 helices; bundle, closed; left-handed twist; 2 crossover connections
4. Superfamily: 4-helical cytokines  
there are two different topoisomers of this fold with different entanglements of the two crossover connections
5. Family: Short-chain cytokines

Protein Domains:

1. Erythropoietin  
long chain cytokine with a short-chain cytokine topology
  1. Human (Homo sapiens) (3)
2. Granulocyte-macrophage colony-stimulating factor (GM-CSF)
  1. Human (Homo sapiens) (2)
3. Interleukin-4 (IL-4)
  1. Human (Homo sapiens) (13)
4. Interleukin-5  
intertwined dimer
  1. Human (Homo sapiens) (1)
5. Macrophage colony-stimulating factor (M-CSF)  
forms dimer similar to the Flt3 ligand and SCF dimers
  1. Human (Homo sapiens) (1)

Etc.

**AMGEN**

WASHINGTON

# Protein Structure

## Four levels of protein structure Primary - amino acid sequence

>gi|119526|sp|P01588|EPO\_HUMAN Erythropoietin  
precursor (Epoetin)

PPRLICDSRVLERYLLEAKEAENITTTGCAEHCSLNENITVPDTKVNIFYAW  
KRMEVGQQAVEVWQGLALLSEAVLRGQALLVNSSQPWEPLQLHVDKAVSG  
LRSLTTLLRALGAQKEAISPDAASAAPLRTITADTFRKLF RVYSNFLRG  
KLKLYTGEACRTGDR

**Efficiency Of Signalling Through Cytokine Receptors Depends Critically On Receptor Orientation**, R.S.Syed, S.W.Reid, C.Li, J.C.Cheetham, K.H.Aoki, b.Liu, H.Zhan, T.D.Osslund, A.J.Chirino, J.Zhang, J.Finer-Moore, S.Elliott, K.Sitney, B.A.Katz, B.J.Matthews, J.J.Wendoloski, J.Egrie, R.M.Stroud, Nature, V. 395, 511, 1998.

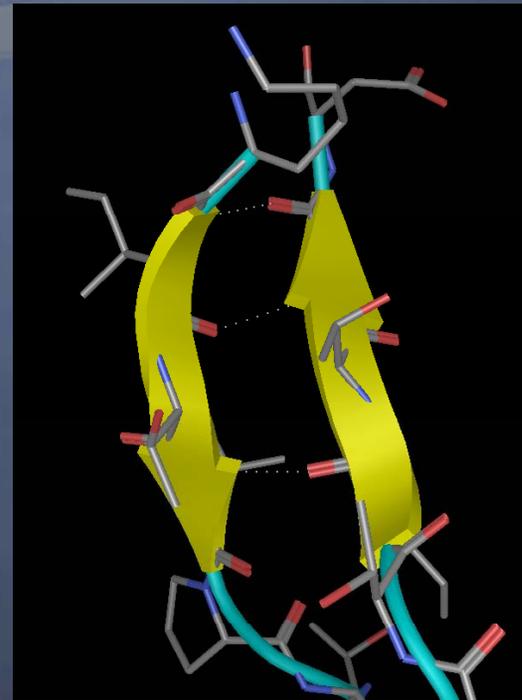
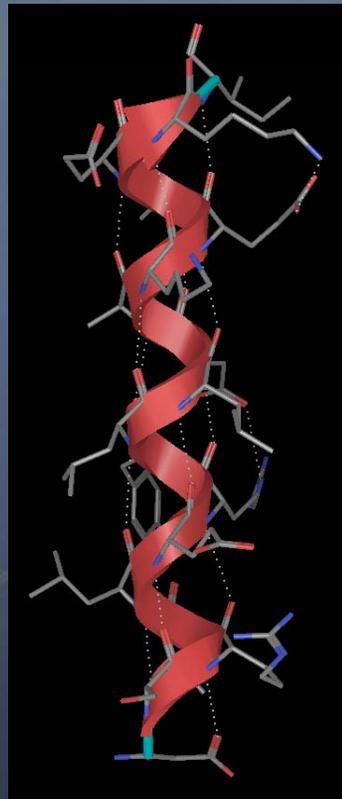
**AMGEN**

WASHINGTON

# Protein Structure

Four levels of protein structure

Secondary - local structure such as  $\alpha$  helices and  $\beta$  strands



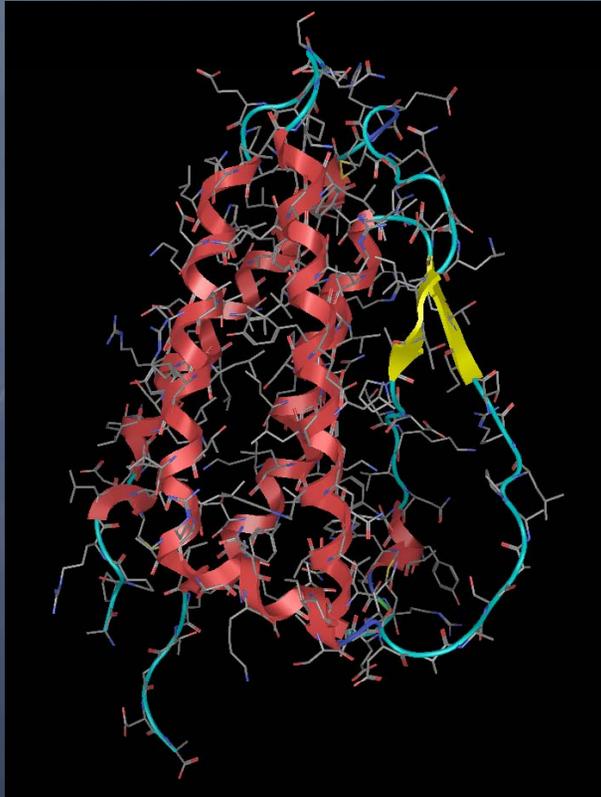
AMGEN

WASHINGTON

# Protein Structure

Four levels of protein structure

Tertiary - packing secondary structure elements into domains



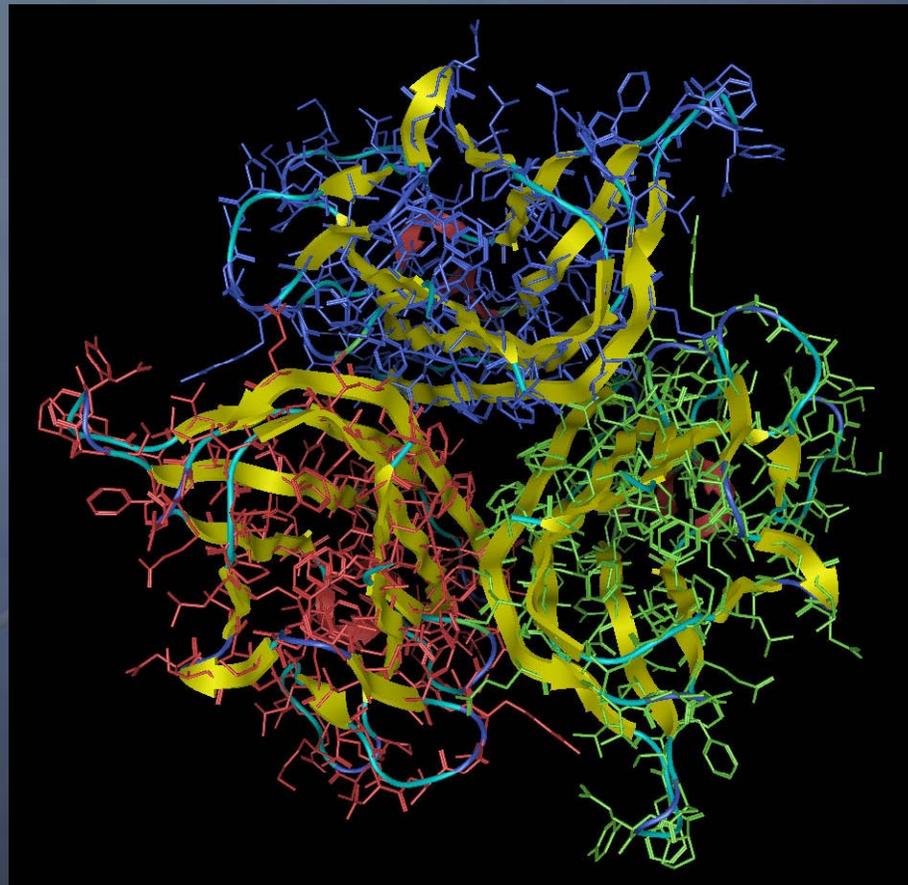
AMGEN

WASHINGTON

# Protein Structure

Four levels of protein structure

Quaternary - multiple chains



**AMGEN**

WASHINGTON

# Experimental Structure

Proteins too small to see

Solid State NMR

Solution NMR

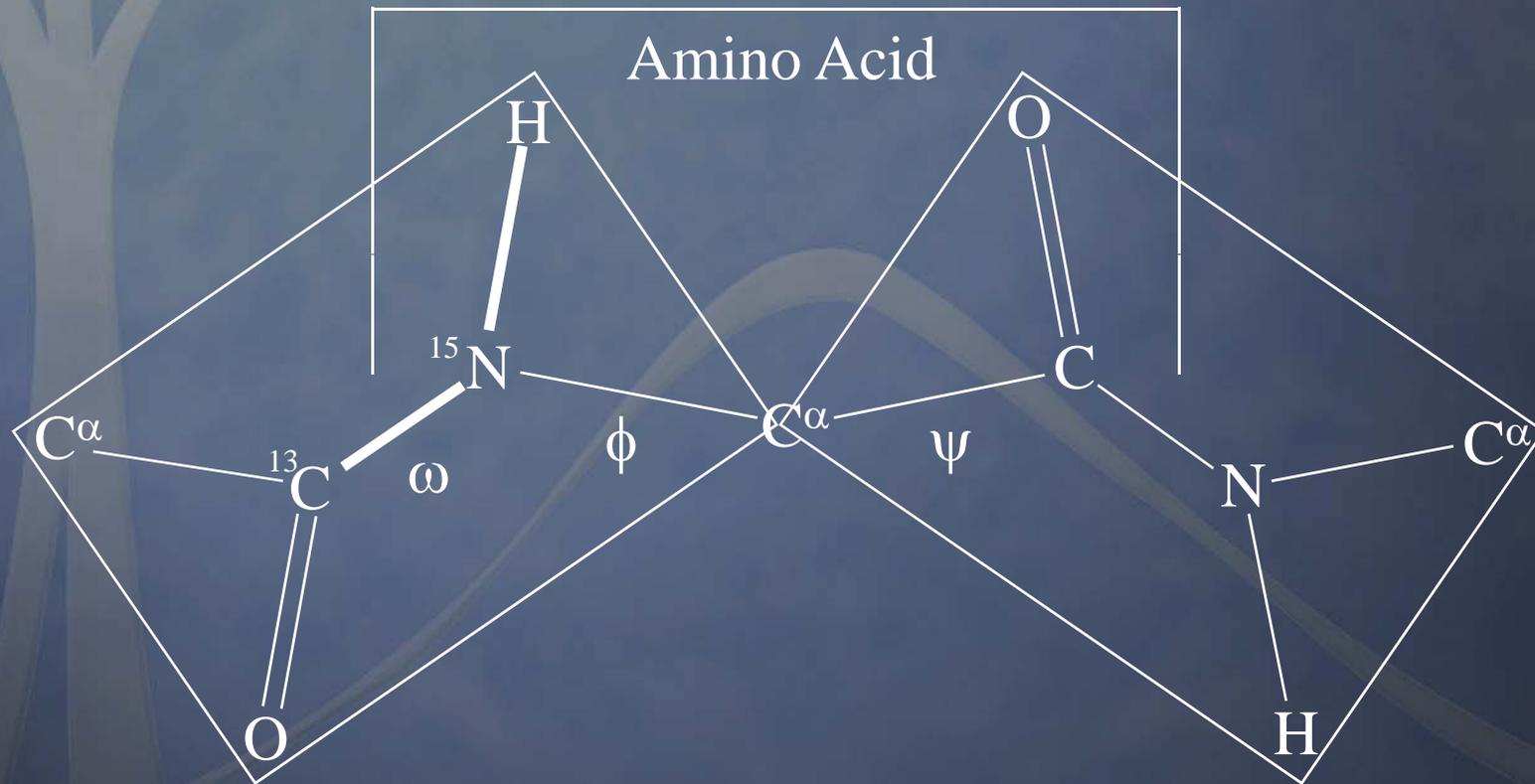
X-Ray Crystallography

**AMGEN**

WASHINGTON

# Solid State NMR

Backbone consists of diplanes

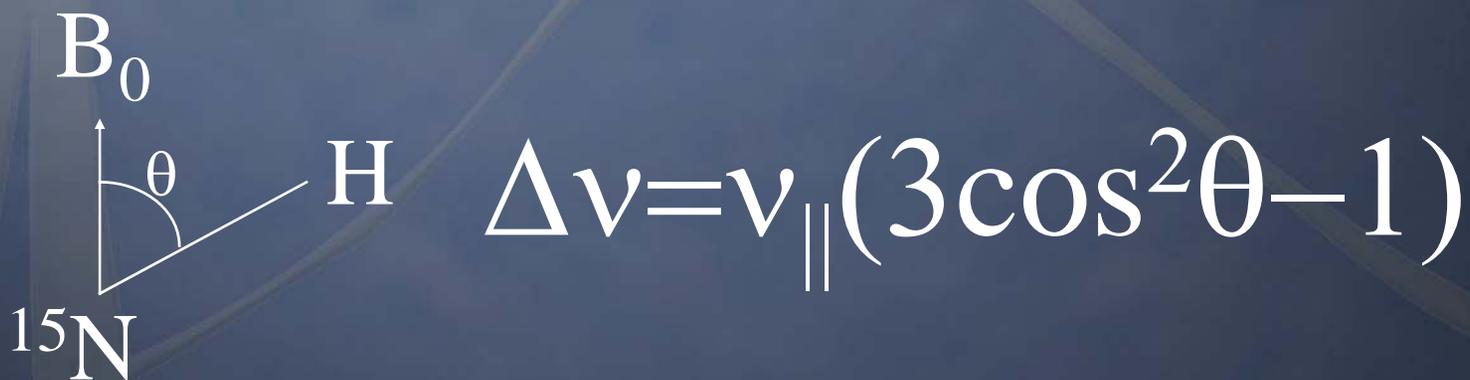


# Solid State NMR

Bond angles measurable to external magnetic field

Two intersecting vectors defines plane orientation

Join planes to determine dihedrals



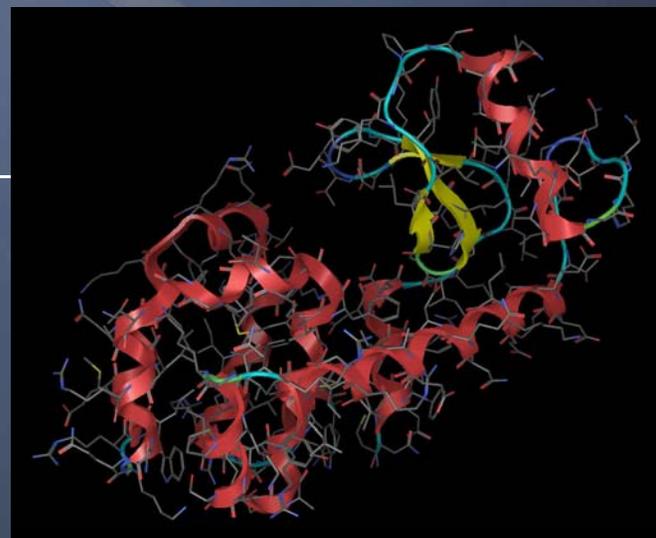
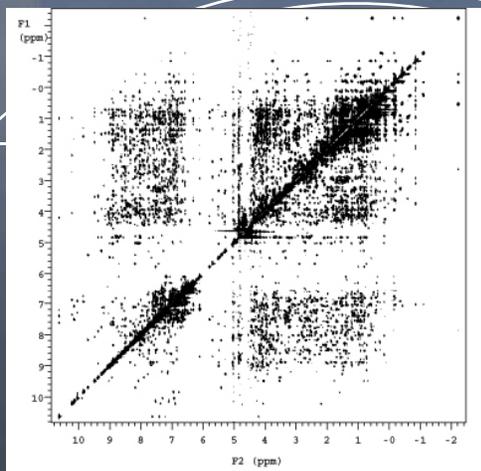
# Solution NMR

Magnetization transfers between nuclei

Distance dependent

Assign measured NOE's to atoms

Fold structure using Distance Geometry



**AMGEN**

WASHINGTON

# X-Ray Crystallography

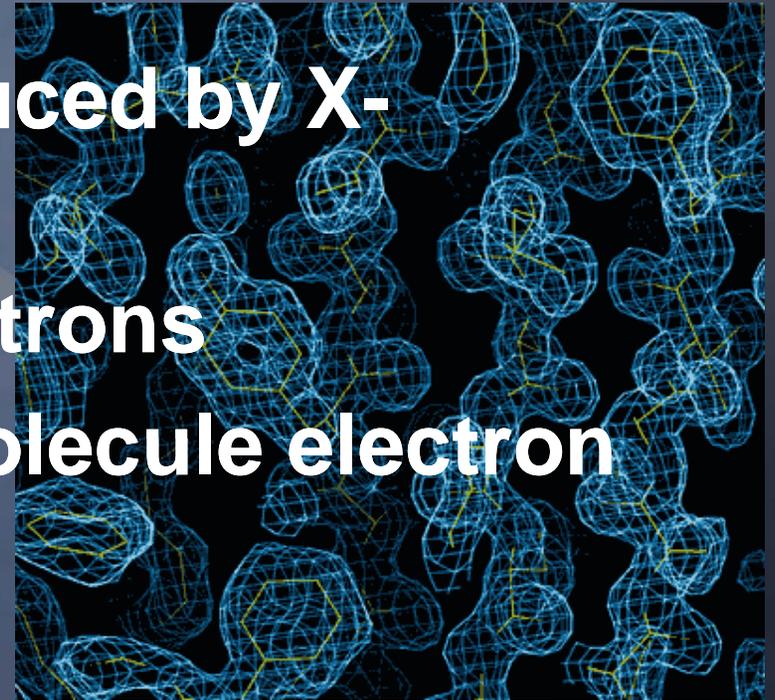
Molecule crystallized, crystals singular,

perfect quality

Diffraction pattern produced by X-irradiation

X-rays diffracted by electrons

Result is 3D image of molecule electron clouds



AMGEN

WASHINGTON

# Homology Modeling

Align sequence with unknown structure  
to sequence with known structure

Extract structural parameters from  
known and apply to unknown

Evaluate, modify alignment, and repeat

Higher homology produces more  
accurate homology model

**AMGEN**

WASHINGTON

# Structure Prediction

Homology modeling is routine with sequence identity  $> 30\%$

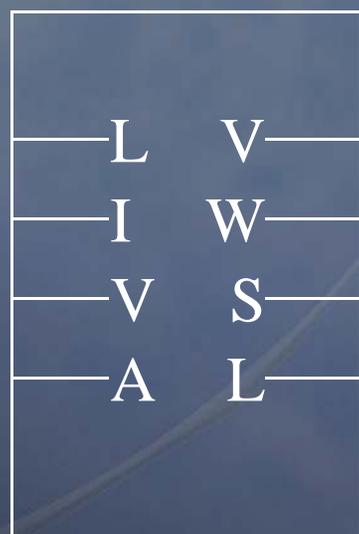
Less than 25% homology is termed the twilight zone and requires other methods

Protein Structure Prediction Using Inverse Folding (Threading)

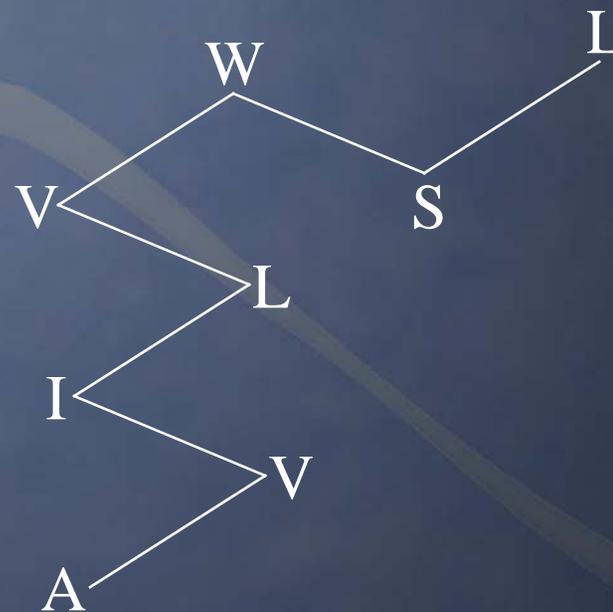
# Threading

“Thread” a protein sequence onto a known structure

Score the threaded fold



Happy



Sad

# GeneFold Threading

Uses a representative library of protein folds and various fitness functions to find the most appropriate fold for a given probe sequence



KPAAHLIGDPSKQNSLLWRANTDRAFLQDGFSLSNNSLLVPTSGIYFVYSQVVVFSGKAYS  
PKATSSPLYLAHEVQLFSSQYPFHVPLLSQKMYVYPGLQEPWLHSMYHGAAFQLTQGDQL  
STHTDGI PHLVLS PSTVFFGAFAL

L.Jaroszewski, L.Rychlewski, B.Zhang and A.Godzik "Fold Predictions by a Hierarchy of Sequence, Threading and Modeling Methods" Protein Science 7:1431-1440 (1998).

AMGEN

WASHINGTON

# GeneFold Threading

Describes each template protein in terms of:

Sequence

Burial pattern of residues

Local main chain conformation

Secondary structure classification



KPAAHLIGDPSKQNSLLWRANT  
DRAFLQDGFSLSNNSLLVPTSG  
IYFVYSQVVFSGKAYS



AMGEN

WASHINGTON

# GeneFold Threading

Structure database based on PDB

Clustered by 50% sequence identity

Theoretical, long (>900) and short (<40) structures removed

1500 Clusters - highest resolution structure chosen as representative (if no x-ray, choose NMR - grr)

**AMGEN**

WASHINGTON

# GeneFold Threading

Scores a target sequence using:

Sequence-sequence: No structural information

Sequence-structure: Pseudo-energy of a single residue mounted in the template structural environment

Structure-structure: Comparison between predicted and actual secondary structure

# GeneFold Threading

## Three scoring methods

Sequence similarity: sequence term only

Hybrid sequence/structure similarity:  
sequence, local conformation and burial

Full hybrid: Sequence, secondary  
structure, local conformation and burial

# GeneFold Threading

No one method produces a reliable prediction, but different methods give consistently correct answers

## Jury Prediction

Two methods agree or

One of the three has a high reliability

# GeneFold Threading

## GeneFold Scores

A given probe is aligned with every template and scored

P-value is calculated for alignment ensemble using distribution of scores

The inverse of the P-value is reported

This process is repeated independently for the three methods

# Mining the Genome

Database of all gene predictions  
translated to protein sequences

Calculate GeneFold scores for each  
sequence

Relate interesting families using  
known proteins

Search by family

**AMGEN**

WASHINGTON

# Mining the Genome

## An example: Mining the Family of Interleukins

### Celera Genefold Data

Celera human r26b and mouse r12 Otto predictions and GeneFold 6.7  
[instructions](#)

Enter a Celera ID (HCP...):

Human:  Mouse:

Sort by:

**Or, Select a GeneFold family as related to ProtBase (ProtBase Category: Possible GeneFold family):**

•Bear in mind that this is merely an alternate method of choosing the GeneFold family. As such, several ProtBase categories map to the same GeneFold family, and therefore provide an identical list of Celera id's, regardless of belonging to different ProtBase categories. For example, 4BHC:1lki\_CYTOKINE is identical to RTK-CSF:1lki\_CYTOKINE.

•This list is prepared by running all ProtBase proteins classified as known through GeneFold and selecting the strong hits from those runs. The hits are then sorted and the known assignment is associated with its possible GeneFold families. This is merely a help in choosing GeneFold families to mine. The comprehensive list of possible GeneFold families is available below.

•The listed families contain PDB ID's. The first four characters are the PDB ID. The last character is the chain. An underscore indicates that there is a single chain for this ID. You can get details for a PDB ID at <http://www.rcsb.org/pdb/>

FIL:1itn_ BINDING PROTEIN
FIL:2frtE COMPLEX (RECEPTOR/IMMUNOGLOBULIN)
FIL:1ita_ CYTOKINE
FIL:7i1b_ CYTOKINE
FILR:1ic1B CELL ADHESION
FILR:1vscB CELL ADHESION PROTEIN

**AMGEN**

WASHINGTON

# Mining the Genome

## Browse the hits for the selected PDB chain

Celera IDs for which family 'lita\_ CYTOKINE' is possible:

GeneBase info color code:

Known (and source not celera or sanger)  Known and Categorized  Unknown

Contig numbers are relative to the chromosome. Protein numbers are relative to the contig, with the exons ordered, begin being the begin of the first exon, end being the end of the last exon.

Human:  Mouse:

Sort by:  [Method for Sorting This Table](#)

Show only hits where:

family is at least number  and score is at least  (zero ignores these cutoffs).

<a href="#">HCP34318.1</a> Sequence Info Known: IL-1 family GeneBase (IMX189)	Family is number 1 of 15 possible score 999.9, length 277	contig: GA_X54KRE9YM0J chrom: 2 begin: 108313232 end: 109165726 Protein begin: 350388 end: 340348
<a href="#">HCP34322.1</a> Sequence Info Known: IL-1 family GeneBase (IMX115)	Family is number 1 of 15 possible score 999.9, length 298	contig: GA_X54KRE9YM0J chrom: 2 begin: 108313232 end: 109165726 Protein begin: 402705 end: 395685
<a href="#">HCP1628454.1</a> Sequence Info Unknown GeneBase (IMX181783)	Family is number 1 of 15 possible score 163.1, length 251	contig: GA_X54KREBBWRK chrom: 8 begin: 121558207 end: 136784473 Protein begin: 7220385 end: 7250443

AMGEN

WASHINGTON

# Mining the Genome

## Drill in on possible hits

**Possible GeneFold Families for**  
ID: HCP1628454.1  
Description: /len=251 /protein\_uid=101000087703514 /ga\_name=GA\_x54KREBBWRK /ga\_uid=181000067218227 /transcript\_na  
Length: 251  
[Sequence Info](#)  
Unknown  
[GeneBase \(IMX181783\)](#)

<input type="button" value="Run GeneFold"/>	<b>FASTA Sequence:</b> >HCP1628454.1-PC1 /len=251 /protein_uid=101000087703514 /ga_name=GA_x54KREBBWRK /g MLIKINQOKVETEWNLTHTQSAVVRVEEMDTVSLPKAKENKNGIECYAPELSKFTFCIHPSIKGPKLLMYAIV AGVFQDKNTLIFQKCSKMGTSARANSREIPSTFLWLKKSVHRRHLAVVDSYWCMSAGPRSGISAGGKHLPLVP SGAASYLRSVTLIWDLSGILERVHHPKAEAAWGPKYLHAQPAAGTPPCCQEAQGLAHPDPPLAPKYDAQGLKQ KAGPLPTPLVPEDHPSGVLFPRKDSF
<input type="button" value="Blast vs. GeneBase"/>	

### GeneFold Hits:

<b>lita_ CYTOKINE</b> Journal Title: STRUCTURE AND FUNCTION OF INTERLEUKIN-1, BASED ON CRYSTALLOGRAPHIC AND MODELING STUDIES score: <b>163.1</b> , score type: <b>br</b> , hit name: <a href="#">lita_</a>
<b>1volA COMPLEX(TRANSCRIPTION FACTOR/REGN/DNA)</b> Journal Title: CRYSTAL STRUCTURE OF A TFIIB-TBP-TATA-ELEMENT TERNARY COMPLEX score: <b>134.1</b> , score type: <b>br</b> , hit name: <a href="#">1volA</a>
<b>4rhv3 RHINOVIRUS COAT PROTEIN</b> Journal Title: THE USE OF MOLECULAR-*REPLACEMENT PHASES FOR THE REFINEMENT OF THE HUMAN RHINOVIRUS 14 STRUCTURE score: <b>132.8</b> , score type: <b>sq</b> , hit name: <a href="#">4rhv3</a>
<b>ludiE COMPLEX (HYDROLASE/INHIBITOR)</b> Journal Title: NUCLEOTIDE MIMICRY IN THE CRYSTAL STRUCTURE OF THE URACIL-DNA GLYCOSYLASE - URACIL GLYCOSYLASE INHIBITOR PROTEIN COMPLEX score: <b>130.9</b> , score type: <b>br</b> , hit name: <a href="#">ludiE</a>
<b>lita_ CYTOKINE</b> Journal Title: STRUCTURE AND FUNCTION OF INTERLEUKIN-1, BASED ON CRYSTALLOGRAPHIC AND MODELING STUDIES score: <b>129.6</b> , score type: <b>sq</b> , hit name: <a href="#">lita_</a>

**AMGEN**

WASHINGTON

# Mining the Genome

## Verify by viewing full GeneFold run

GeneFold: Display Run ketchemr\_20020520103949.txt  
 (Instructions) (Return to the [main GeneFold page](#)) (Display Printable Version)

seq. (sq)			seq. +loc. +bur. (br)			seq. +loc. +bur. +ss (tt)		
* 1.	<a href="#">4rhv3</a>	P F 132.8 4rhv3 RHINOVIRUS C	* 1.	<a href="#">lita</a>	P F 163.1 lita CYTOKINE (ST	* 1.	<a href="#">ludiE</a>	P F 39.7 ludiE COMPLEX (HYDR
* 2.	<a href="#">lita</a>	P F 129.6 lita CYTOKINE (ST	* 2.	<a href="#">ivolA</a>	P F 134.1 ivolA COMPLEX(TRAN	* 2.	<a href="#">lita</a>	P F 33.8 lita CYTOKINE (STRU
* 3.	<a href="#">ludiE</a>	P F 111.3 ludiE COMPLEX (HYD	* 3.	<a href="#">ludiE</a>	P F 130.9 ludiE COMPLEX (HYD	* 3.	<a href="#">lntyH</a>	P F 25.9 lntyH MONOOXYGENASE
* 4.	<a href="#">lurk</a>	P F 109.5 lurk PLASMINOGEN	* 4.	<a href="#">4eng</a>	P F 77.2 4eng GLYCOSYL HYD	* 4.	<a href="#">lurk</a>	P F 22.9 lurk PLASMINOGEN AC
* 5.	<a href="#">9pap</a>	P F 87.5 9pap HYDROLASE (S	* 5.	<a href="#">2cab</a>	P F 67.3 2cab HYDRO-LYASE	* 5.	<a href="#">lgnbJ</a>	P F 18.8 lgnbJ ACUTE-PHASE PR
* 6.	<a href="#">2cab</a>	P F 76.3 2cab HYDRO-LYASE	* 6.	<a href="#">lthw</a>	P F 61.0 lthw SWEET TASTIN	* 6.	<a href="#">4rhv3</a>	P F 17.6 4rhv3 RHINOVIRUS COA
* 7.	<a href="#">lntyH</a>	P F 69.4 lntyH MONOOXYGENAS	* 7.	<a href="#">lgnbJ</a>	P F 55.5 lgnbJ ACUTE-PHASE	* 7.	<a href="#">lc52</a>	P F 15.6 lc52 ELECTRON TRANS
* 8.	<a href="#">ivolA</a>	P F 62.0 ivolA COMPLEX(TRAN	* 8.	<a href="#">7timB</a>	P F 49.2 7timB INTRAMOLECUL	* 8.	<a href="#">lfxrB</a>	P F 14.8 lfxrB ELECTRON TRANS
* 9.	<a href="#">2hipH</a>	P F 60.5 2hipH COMPLEX (AMT	* 9.	<a href="#">lsvcP</a>	P F 46.1 lsvcP COMPLEX (TRA	* 9.	<a href="#">lkst</a>	P F 14.4 lkst AGGREGATION IM
* 10.	<a href="#">lppo</a>	P F 59.6 lppo HYDROLASE (TH	* 10.	<a href="#">lmri</a>	P F 44.2 lmri RIBOSOME-LNA	* 10.	<a href="#">lsebF</a>	P F 14.1 lsebF COMPLEX (MHC I
* 11.	<a href="#">lgnbJ</a>	P F 57.7 lgnbJ ACUTE-PHASE	* 11.	<a href="#">lsebF</a>	P F 42.2 lsebF COMPLEX (MHC	* 11.	<a href="#">7timB</a>	P F 13.5 7timB INTRAMOLECULAR
* 12.	<a href="#">lsebF</a>	P F 55.4 lsebF COMPLEX (MHC	* 12.	<a href="#">ljud</a>	P F 42.0 ljud DEHALOGENASE	* 12.	<a href="#">liceA</a>	P F 12.7 liceA CYTOKINE (STRU
* 13.	<a href="#">6fabH</a>	P F 53.3 6fabH IMMUNOGLOBUL	* 13.	<a href="#">lpoiD</a>	P F 38.7 lpoiD TRANSFERASE	* 13.	<a href="#">ivolA</a>	P F 12.4 ivolA COMPLEX(TRANSC
* 14.	<a href="#">3drcB</a>	P F 52.4 3drcB OXIDOREDUCTA	* 14.	<a href="#">4jdwA</a>	P F 38.6 4jdwA TRANSFERASE	* 14.	<a href="#">lme1B</a>	P F 12.3 lme1B COMPLEX (ANTIB
* 15.	<a href="#">6fabI</a>	P F 52.0 6fabI IMMUNOGLOBUL	* 15.	<a href="#">law9</a>	P F 35.4 law9 TRANSFERASE	* 15.	<a href="#">3dfr</a>	P F 12.1 3dfr OXIDO-REDUCTAS
* 16.	<a href="#">3sdpB</a>	P F 44.7 3sdpB OXIDOREDUCTA	* 16.	<a href="#">2fvlw</a>	P F 35.2 2fvlw IMMUNOGLOBUL	* 16.	<a href="#">ljvr</a>	P F 12.0 ljvr MATRIX PROTEIN
* 17.	<a href="#">lan3C</a>	P F 40.6 lan3C COMPLEX (HORM	* 17.	<a href="#">8fabD</a>	P F 34.9 8fabD IMMUNOGLOBUL	* 17.	<a href="#">lvcqB</a>	P F 11.7 lvcqB NUCLEOCAPSID P
* 18.	<a href="#">lmpaH</a>	P F 36.6 lmpaH COMPLEX (IMM	* 18.	<a href="#">larc</a>	P F 34.7 larc HYDROLASE (SE	* 18.	<a href="#">lthw</a>	P F 11.6 lthw SWEET TASTING
* 19.	<a href="#">lcv8</a>	P F 34.9 lcv8 CYSTEINE PRO	* 19.	<a href="#">lcrA</a>	P F 34.3 lcrA COMPLEX (DNA	* 19.	<a href="#">lB5m</a>	P F 11.6 lB5m ELECTRON TRANS
* 20.	<a href="#">ljud</a>	P F 34.1 ljud DEHALOGENASE	* 20.	<a href="#">lp38</a>	P F 34.2 lp38 TRANSFERASE	* 20.	<a href="#">6fabH</a>	P F 11.6 6fabH IMMUNOGLOBULIN
* 21.	<a href="#">lhnf</a>	P F 33.8 lhnf T LYMPHOCYTE	* 21.	<a href="#">2bbvB</a>	P F 34.1 2bbvB COMPLEX(VIRU	* 21.	<a href="#">lnfa</a>	P F 11.5 lnfa TRANSCRIPTION
* 22.	<a href="#">4eng</a>	P F 33.3 4eng GLYCOSYL HYD	* 22.	<a href="#">leaf</a>	P F 33.8 leaf DIHYDROLIPOA	* 22.	<a href="#">2cab</a>	P F 11.5 2cab HYDRO-LYASE
* 23.	<a href="#">lwdcB</a>	P F 31.5 lwdcB MUSCLE PROTE	* 23.	<a href="#">lvcqB</a>	P F 32.9 lvcqB NUCLEOCAPSID	* 23.	<a href="#">lcd8</a>	P F 11.1 lcd8 SURFACE GLYCOP
* 24.	<a href="#">lcnv</a>	P F 31.3 lcnv SEED PROTEIN	* 24.	<a href="#">lmcD</a>	P F 32.4 lmcD HISTOCOMPATI	* 24.	<a href="#">lncpC</a>	P F 11.1 lncpC COMPLEX (HORMO
* 25.	<a href="#">lqc1H</a>	P F 31.0 lqc1H COMPLEX (HIV	* 25.	<a href="#">lako</a>	P F 32.4 lako NUCLEASE	* 25.	<a href="#">lyaiC</a>	P F 11.0 lyaiC OXIDOREDUCTASE
* 26.	<a href="#">3hfmH</a>	P F 27.3 3hfmH COMPLEX(ANTI	* 26.	<a href="#">2hvm</a>	P F 31.4 2hvm HYDROLASE (T	* 26.	<a href="#">ltiv</a>	P F 10.9 ltiv TRANSCRIPTION
* 27.	<a href="#">lxl</a>	P F 26.1 lxl APOPTOSIS (X	* 27.	<a href="#">2acu</a>	P F 31.3 2acu OXIDOREDUCTA	* 27.	<a href="#">lhunB</a>	P F 10.5 lhunB CYTOKINE(CHEMO
* 28.	<a href="#">8fabD</a>	P F 25.8 8fabD IMMUNOGLOBUL	* 28.	<a href="#">lnbaD</a>	P F 30.8 lnbaD HYDROLASE (IN	* 28.	<a href="#">lar1D</a>	P F 10.3 lar1D COMPLEX (OXIDO
* 29.	<a href="#">ltam</a>	P F 25.6 ltam MATRIX PROTE	* 29.	<a href="#">lbbmB</a>	P F 29.0 lbbmB COMPLEX (ENL	* 29.	<a href="#">4vqcC</a>	P F 10.3 4vqcC SERINE PROTEAS
* 30.	<a href="#">lvcqB</a>	P F 23.9 lvcqB NUCLEOCAPSID	* 30.	<a href="#">6fabI</a>	P F 28.7 6fabI IMMUNOGLOBUL	* 30.	<a href="#">lrtoB</a>	P F 10.1 lrtoB CHEMOKINE (PRO
* 31.	<a href="#">lppB</a>	P F 23.0 lppB COMPLEX (HOM	* 31.	<a href="#">lfrfS</a>	P F 28.7 lfrfS NT-EE HYDROG	* 31.	<a href="#">6rbn</a>	P F 10.1 6rbn NUCLEOTIDE-RIN

Alignment of lita\_ method br, run gf  
 50.6% of the query is aligned, 84.1% of the template is aligned

Primary sequence:

query: **MLIK** **ML** **QK** **YET** **EW** **NL** **TE** **HT** **Q** **S** **A** **V** **R** **R** **E** **E** **M** **D** **T** **V** **S** **L** **P** **K** **A** **K** **E** **N** **K** **N** **G** **I** **E** **C** **Y** **A** **P** **E** **L** **S** **K** **---** **F** **T** **P** **C** **I** **H** **P** **S** **---** **I** **K** **G** **P** **K** **L** **M** **A** **T** **A** **G** **V** **F** **Q** **D** **K** **N** **T** **L** **I** **F** **Q** **K** **S** **K** **M** **G** **T** **S** **A** **R** **A** **N** **S** **R** **E** **I** **P** **S** **T** **F** **L** **W**  
 lita\_ : **---** **N** **V** **K** **Y** **N** **F** **M** **R** **---** **L** **I** **K** **Y** **E** **F** **I** **L** **N** **D** **A** **L** **N** **Q** **---** **S** **I** **I** **R** **A** **N** **D** **Q** **Y** **---** **L** **T** **A** **A** **A** **L** **H** **N** **L** **D** **E** **A** **V** **K** **F** **D** **M** **G** **A** **Y** **S** **S** **K** **D** **A** **K** **I** **T** **V** **I** **L** **R** **I** **S** **K** **T** **Q** **L** **V** **T** **A** **Q** **---** **D** **E** **D** **Q** **P** **V** **L** **L** **K** **H** **M** **P** **E** **I** **P** **K** **T** **I** **G** **S** **E** **T** **N** **L** **---** **L** **F** **F** **W**

Secondary structure:

query: **EEEE** **H** **---** **EEEE** **---** **GGGGGG** **---** **EEEE** **---** **EEEE** **---** **TT** **-----** **EEEE** **---** **EEEE** **---** **EEEE** **---** **EEEE** **---** **EE** **---** **GGGGG** **---** **EEEE**  
 lita\_ : **EEEE** **EEEEEEEE** **---** **EEEE** **---** **EEEE** **---** **TT** **-----** **EEEE** **---** **EEEE** **---** **EEEE** **---** **EEEE** **---** **EE** **---** **GGGGG** **---** **EEEE**

## Some Results

Gene mining by remote homology detection has been very successful

Identified several novel human cytokines

Verification lead to discovery of further novel cytokines

**AMGEN**

WASHINGTON

## Some Problems

**MANY false positives - many hits to wade through**

**Requires expert in particular family to identify true positives**

**Genes must be verified experimentally**

**Some folds hard to score - TNFR's**

**Not all folds represented**

## **Future Work**

**Utilize advances in remote homology detection**

**As structure representatives grow, so will ability of remote homology detection**

**Utilize fast, automated methods for assigning structure family**

# Fast Threading Data Analysis

To mine threading data, we need programs that:

Repeat interpretation of threading output consistently and quickly

We can train to recognize different folds in the output

Aid protein structure experts by applying similar logic

**AMGEN**

WASHINGTON

# Fast Threading Data Analysis

We chose:

The support vector machine algorithm

Quick to train, quick to give answers

Generates score which can be used as  
measure of confidence in answer  
generated

**AMGEN**

WASHINGTON

# How SVM's Work

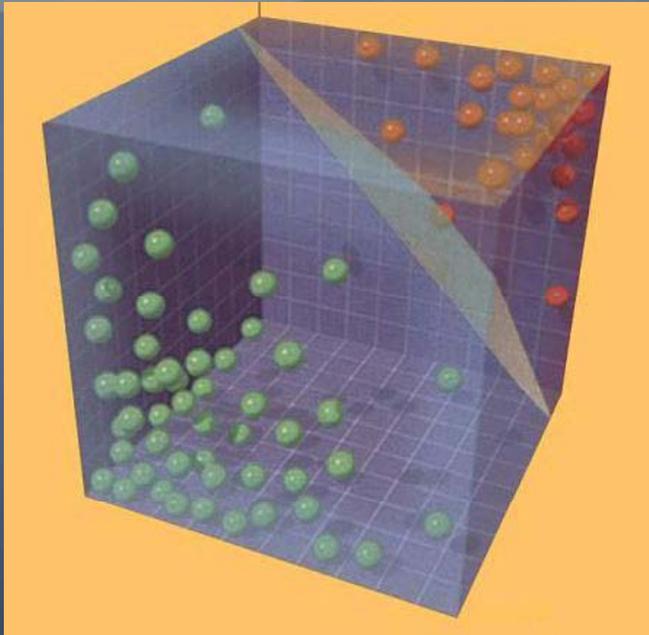


Figure from cover of book: Introduction to Support Vector Machines. Christianni and Shaw-Taylor

**SVM work by Paul Mc Donagh,  
Amgen Inc.**

**SVM takes 'positive' and 'negative' fold examples**

**Red = positive, Green = negative**

**Uses function (kernel function) to plot data to different type of space**

**Graphic shows 3-D linear kernel space**

**Threading uses 1823-D spherical space**

**Regression techniques fit a plane**

**Vectors from points 'support' the plane**

**Term coined - support vector machine**

**If fall on red side of plane - new member of the fold**

**Distance from plane gives measure of confidence in prediction**

**AMGEN**

WASHINGTON

# Support Vector Machines

Trained by scientists

Success primarily depends on  
scientific input to training set

Scientist finds members of fold -  
positive training set

Scientist identifies all other folds -  
negative training set

**AMGEN**

WASHINGTON

# Support Vector Machines

Threading algorithm run on unknowns

1823\*3 data points for each protein in a set

Support vector machines find which of the 1823\*3 points and values carry the most predictive power

Early results have been very promising

**AMGEN**

WASHINGTON

## **Future Work**

**Genome sequenced, but still a LONG way to go for function**

**Structure homology methods valuable in identifying unknown sequences**

**Many structure families not represented**

**Need better remote homology detection methods**

**Need fast, automated methods**

**AMGEN**

WASHINGTON

# Acknowledgments

David Rider - DB and servlets for dealing with data

Paul Mc Donagh - SVMs

Bob DuBose - Supreme Leader

Amgen Washington Scientists - Mining the data

**AMGEN**

WASHINGTON