Mining the Human Genome Using Protein Structure Homology

Randal R. Ketchem, Ph.D.

Immunex Corporation

Introduction

- Need for gene mining
- Scale of problem
- Protein structure
- Structure prediction
- Mining the genome
- Some results
- Some problems
- Future work

```
TCTCGAGGGCCACGCGTTTAAACGTCGAGGTACCTATCCCGGGCCGCCAC
CATGGCTACAGGCTCCCGGACGTCCCTGCTCCTGGCTTTTGGCCTGCTCT
GCCTGCCCTGGCTTCAAGAGGGCAGTGCAACTAGTTCTGACCGTATGAAA
CAGATAGAGGATAAGATCGAAGAGATCCTAAGTAAGATTTATCATATAGA
GAATGAAATCGCCCGTATCAAAAAGCTGATTGGCGAGCGGACTAGATCTA
GTTTGGGGAGCCGGGCATCGCTGTCCGCCCAGGAGCCTGCCCAGGAGGAG
CTGGTGGCAGAGGAGGACCAGGACCCGTCGGAACTGAATCCCCAGACAGA
AGAAAGCCAGGATCCTGCGCCTTTCCTGAACCGACTAGTTCGGCCTCGCA
GAAGTGCACCTAAAGGCCGGAAAACACGGGCTCGAAGAGCGATCGCAGCC
CATTATGAAGTTCATCCACGACCTGGACAGGACGGAGCGCAGGCAGGTGT
GGACGGGACAGTGAGTGGCTGGGAGGAAGCCAGAATCAACAGCTCCAGCC
CTCTGCGCTACAACCGCCAGATCGGGGAGTTTATAGTCACCCGGGCTGGG
CTCTACTACCTGTACTGTCAGGTGCACTTTGATGAGGGGAAGGCTGTCTA
CCTGAAGCTGGACTTGCTGGTGGATGGTGTGCTGGCCCTGCGCTGCCTGG
AGGAATTCTCAGCCACTGCGGCCAGTTCCCTCGGGCCCCAGCTCCGCCTC
TGCCAGGTGTCTGGGCTGTTGGCCCTGCGGCCAGGGTCCTCCCTGCGGAT
CCGCACCCTCCCCTGGGCCCATCTCAAGGCTGCCCCCTTCCTCACCTACT
TCGGACTCTTCCAGGTTCACTGAGCGGCCGCGGATCTGTTTAAACTAG

MATGSRTSLLLAFGLLCLPWLQEGSATSSDRMKQIEDKIEEILSKIYHIE
NEIARIKKLIGERTRSSLGSRASLSAQEPAQEELVAEEDQDPSELNPQTE
ESQDPAPFLNRLVRPRRSAPKGRKTRARRAIAAHYEVHPRPGQDGAQAGV
DGTVSGWEEARINSSSPLRYNRQIGEFIVTRAGLYYLYCQVHFDEGKAVY
LKLDLLVDGVLALRCLEEFSATAASSLGPQLRLCQVSGLLALRPGSSLRI
RTLPWAHLKAAPFLTYFGLFQVH
```

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

Need For Gene Mining

- Experimental methods too slow for complete classification
- Computational methods for elucidating function needed
- Weeks or months, around $100K, to solve single, globular structure

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

immunex®

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

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 entanglments 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.

Protein Structure

Four levels of protein structure

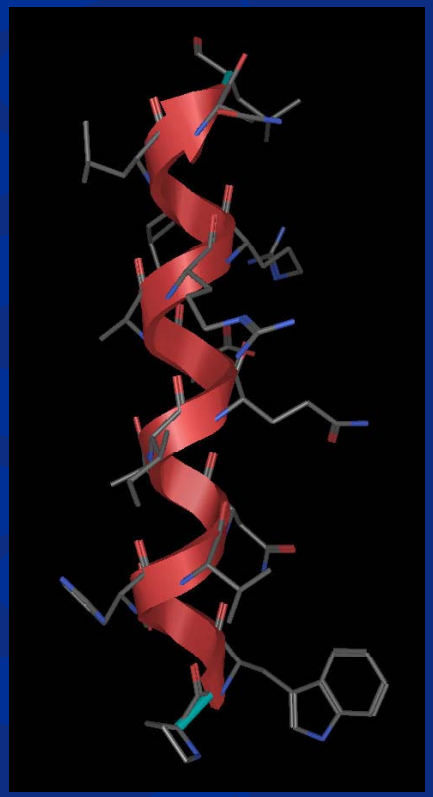
Primary - amino acid sequence

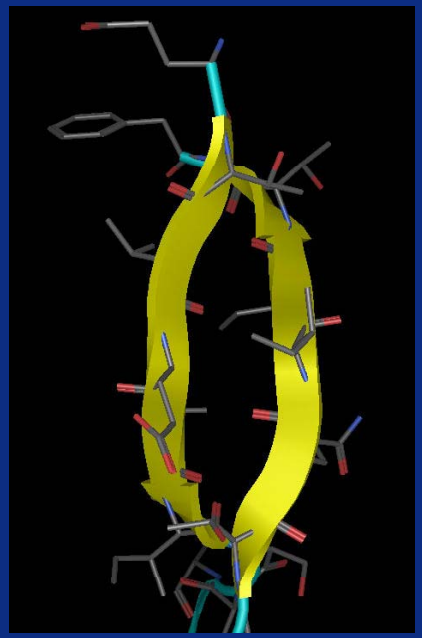
```
>1csgA
SPSPSTQPWEHVNAIQEARRLLNLSRDTAAEMNETVEVISEMFDLQEPTC
LQTRLELYKQGLRGSLTKLKGPLTMMASHYKQHCPPTPETSCATQIITFE
SFKENLKDFLLVIPFDCWEP

GRANULOCYTE-MACROPHAGE COLONY-STIMULATING FACTOR (GM-CSF) HUMAN
(HOMO SAPIENS) RECOMBINANT FORM EXPRESSED IN (ESCHERICHIA COLI)
M.R.WALTER,W.J.COOK,S.E.EALICK
```

Protein Structure

Four levels of protein structure

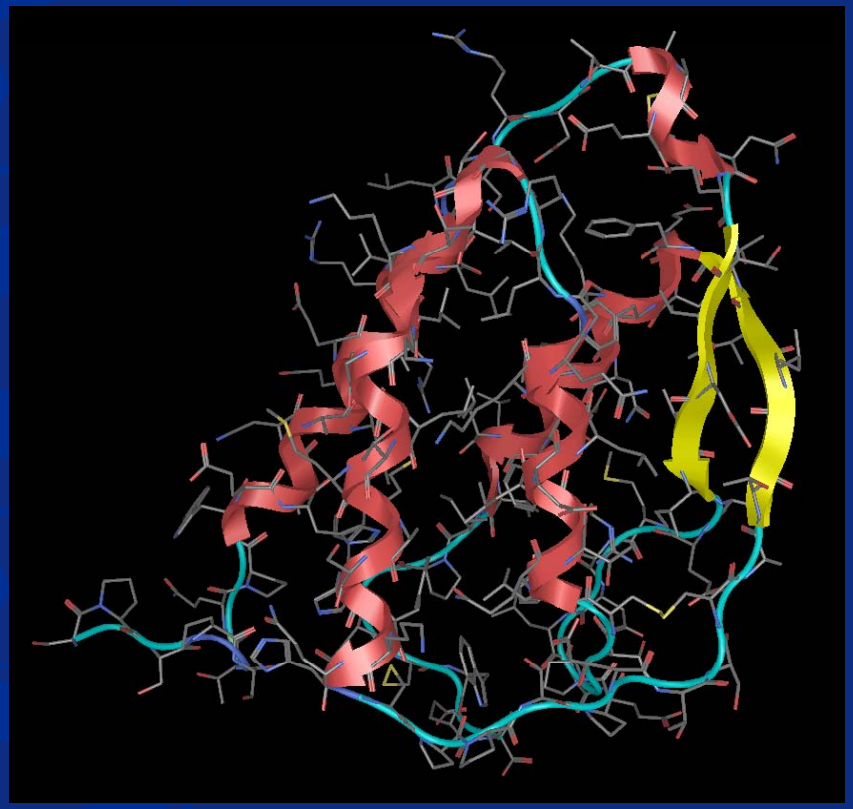
Secondary - local structure such as α helices and β strands





Protein Structure

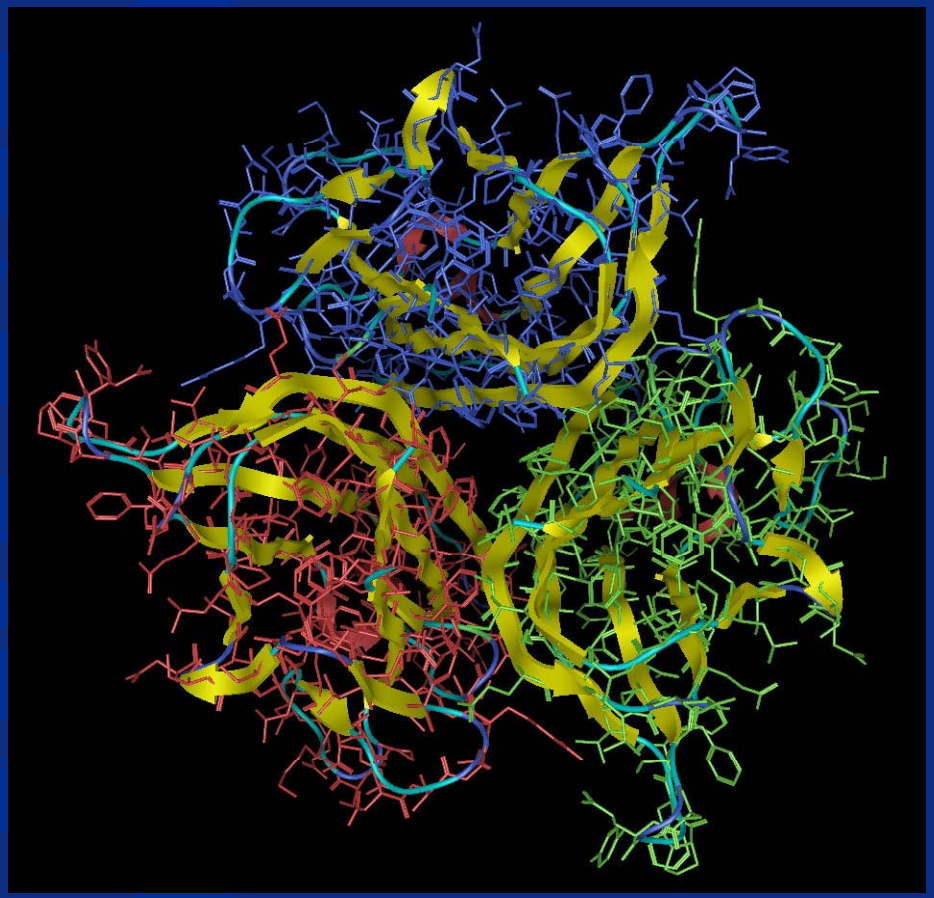
- Four levels of protein structure
- Tertiary - packing secondary structure elements into domains



Protein Structure

Four levels of protein structure

Quaternary - multiple chains



Experimental Structure

Proteins too small to see

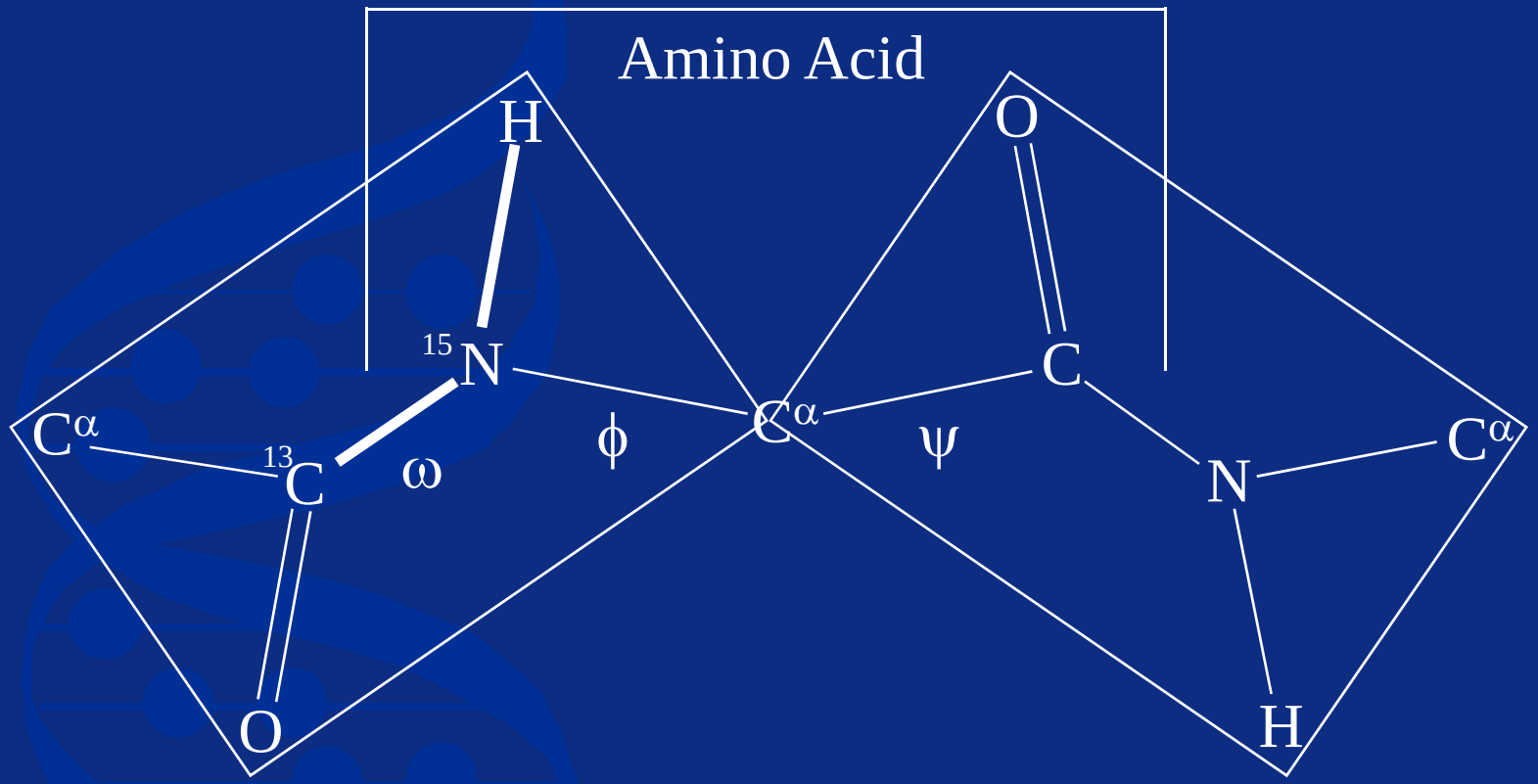
Solid State NMR

Solution NMR

X-Ray Crystallography

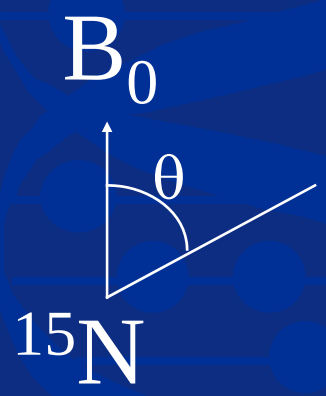
Solid State NMR

Backbone consists of diplanes



Solid State NMR

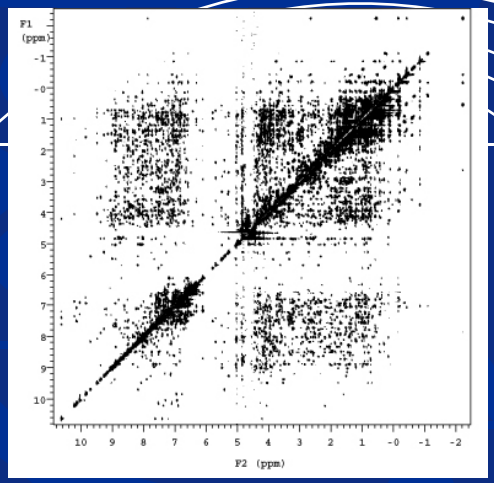
- Bond angles measurable to external magnetic field
- Two intersecting vectors defines plane orientation
- Join planes to determine dihedrals



$$\Delta\nu = \nu_{\parallel}(3\cos^2\theta - 1)$$

Solution NMR

- Magnetization transfers between nuclei
- Distance dependent
- Assign measured NOE's to atoms
- Fold structure using Distance Geometry



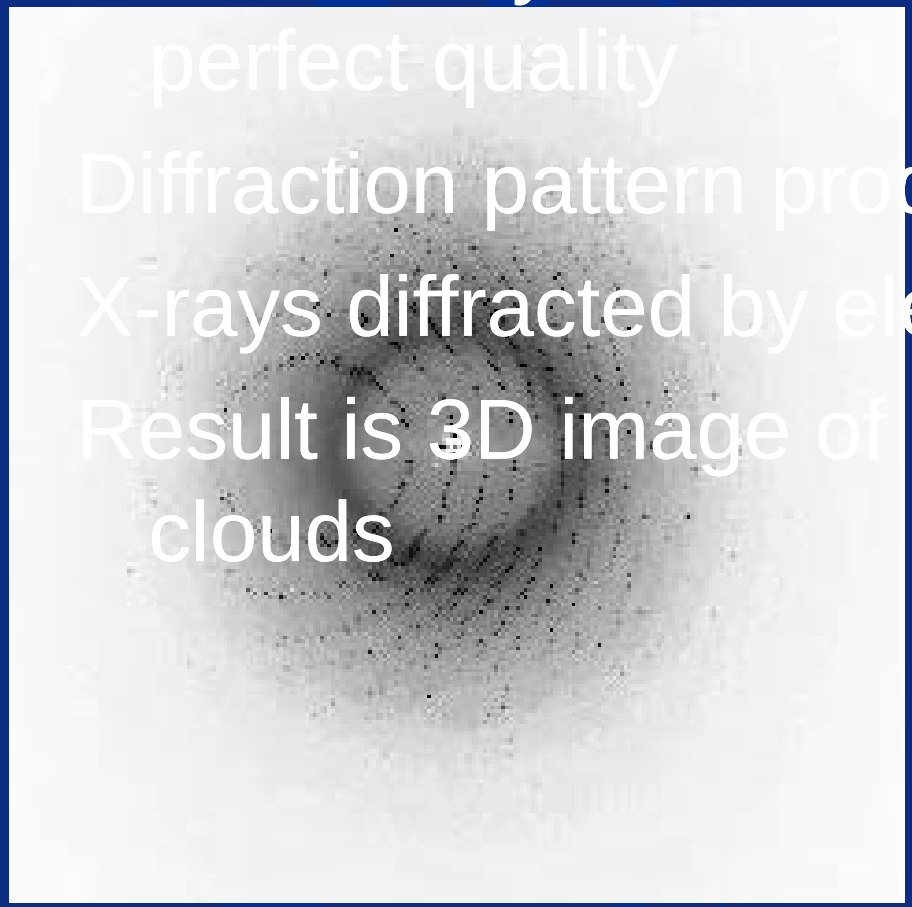
X-Ray Crystallography

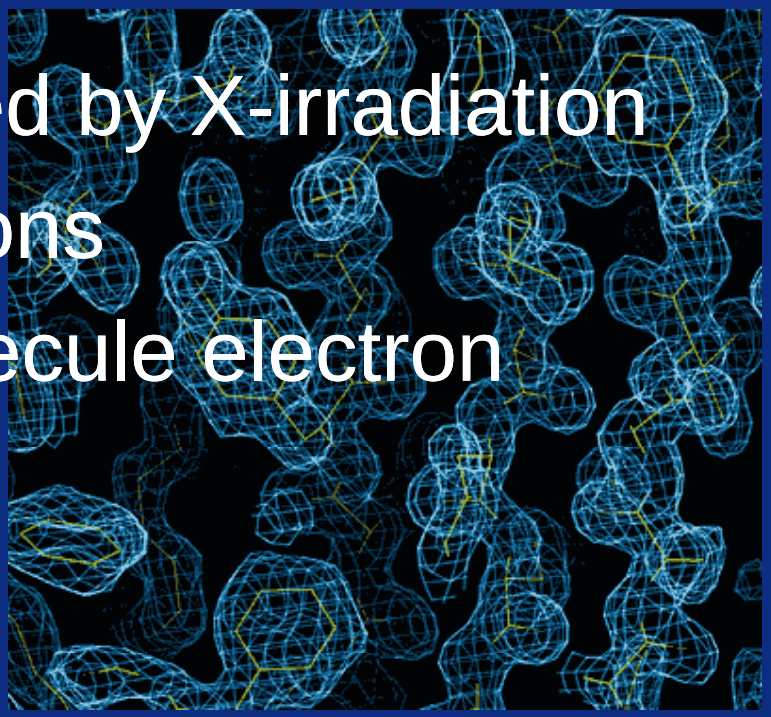
Molecule crystallized, crystals singular, perfect quality

Diffraction pattern produced by X-irradiation

X-rays diffracted by electrons

Result is 3D image of molecule electron clouds





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

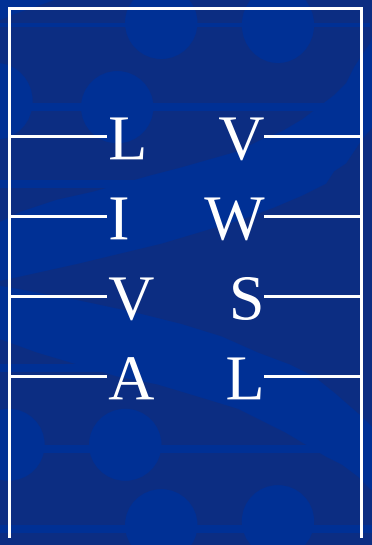
Structure Prediction

- 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)

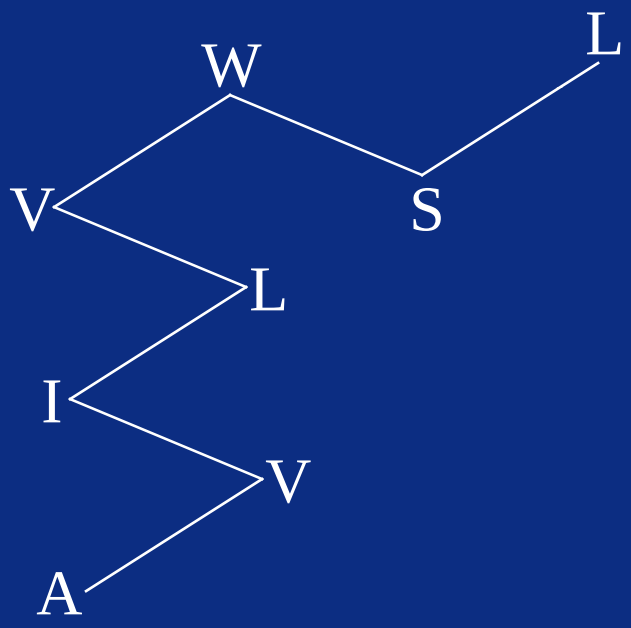
Threading

"Thread" a protein sequence onto a known structure

Score the threaded fold



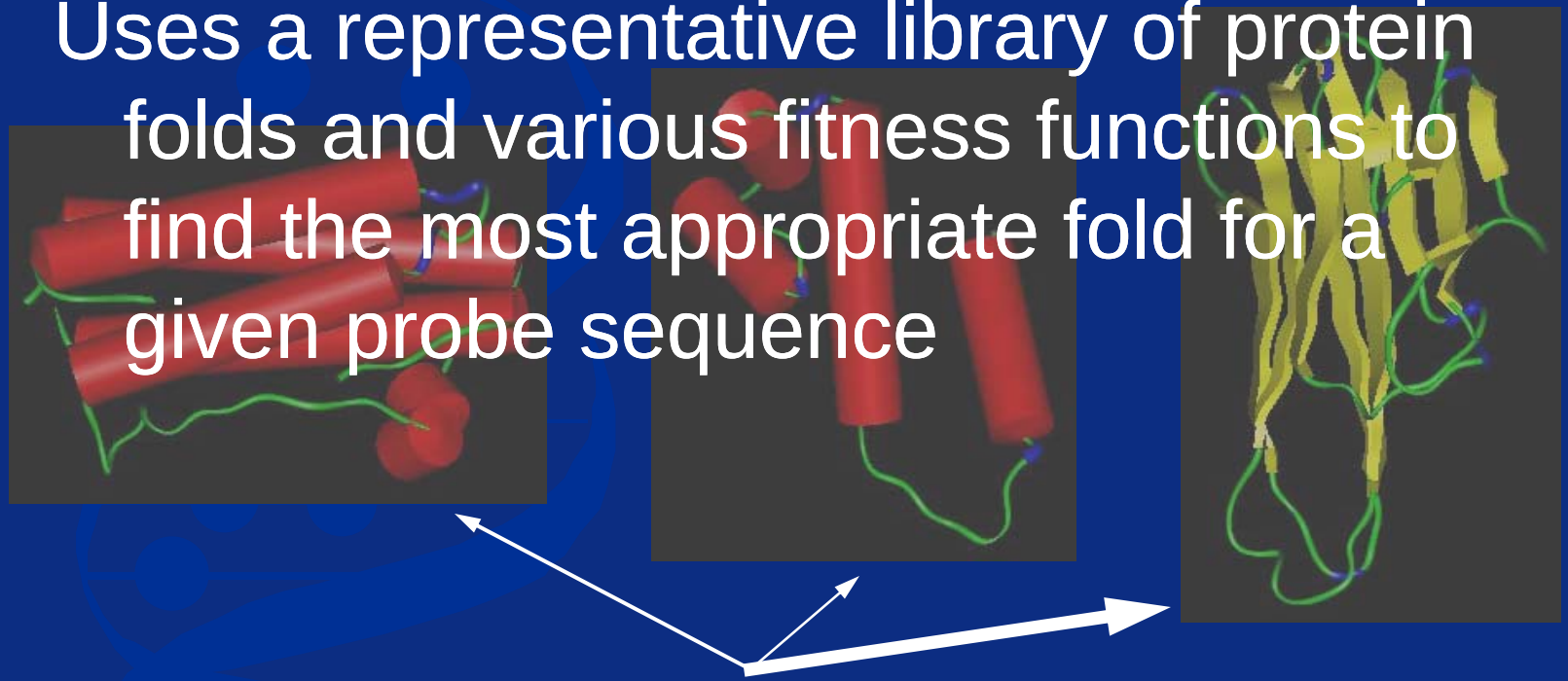
Happy



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



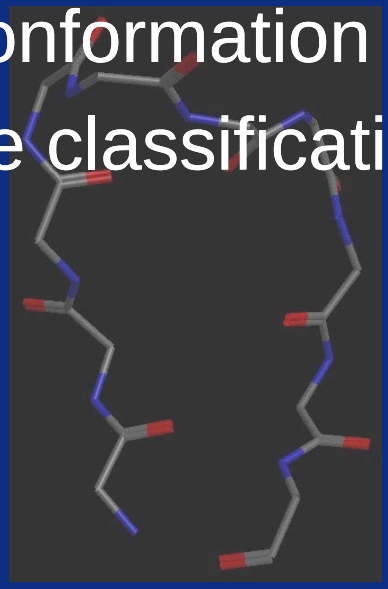
KPAAHLIGDPSKQNSLLWRANTDRAFLQDGFSLSNNSLLVPTSGIYFVYSQVVFSGKAYS
PKATSSPLYLAHEVQLFSSQYPFHVPLLSSQKMVYPGLQEPWLHSMYHGAAFQLTQGDQL
STHTDGIPHLVLSPSTVFFGAFAL

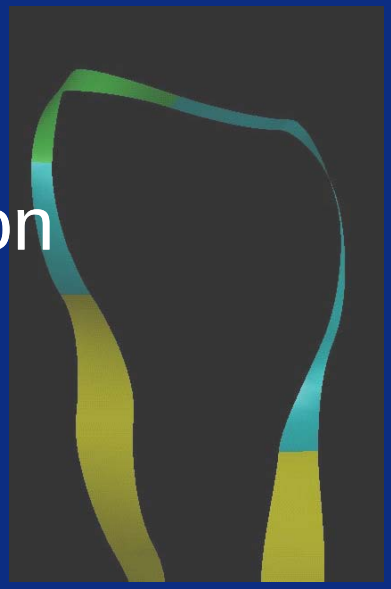
L.Jaroszewski, L.Rychlewski, B.Zhang and A.Godzik "Fold Predictions by a Hierarchy of Sequence, Threading and Modeling Methods" Protein Science 7:1431-1440 (1998).

GeneFold Threading

Describes each template protein in terms of:

- Sequence
- Burial pattern of residues
- Local main chain conformation
- Secondary structure classification





KPAAHLIGDPSKQNSLLWRANT
DRAFLQDGFSLSNNSLLVPTSG
IYFVYSQVVFSGKAYS

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)

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

immunex®

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...):

Submit ID | Reset Form

Human: ☑ Mouse: ☐

Sort by: score

Or, Select a GeneFold family as related to ProtBase (ProtBase Catagory: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.resb.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

immunex®

Mining the Genome

Browse the hits for the selected PDB chain

Celera IDs for which family '1ita_ 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: score Method for Sorting This Table

Show only hits where:
family is at least number 0 and score is at least 0 (zero ignores these cutoffs).
Submit Reset Form

HCP34318.1 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
HCP34322.1 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
HCP1628454.1 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

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)

Run GeneFold

Blast vs. GeneBase

FASTA Sequence:

```
>HCP1628454.1-PC1 /len=251 /protein_uid=101000087703514 /ga_name=GA_x54KREBBWRK /g
MLIKINQQKVETEWNLTEHITQSAYVRVEEMDTVSLPKAKENKNGIECYAPELSKFTFCIHPSIKGPKLLMYAIV
AGVFQDKNTLIFQKCSKMGTSARANSREIPSTFLWLKKSVHRRLHLAVVDSYWCCMSAGPRSGISAGGKHLPLVP
SGAASYLRSVTLIWDLSGILERVHHPSKAEAAWGPKYLHAQPAAGTPPCCQEAQGLAHPDPPLAPKYDAQGLKDQ
KAGPLPTPLVPEDHPSGVLFPRKDSP
```

GeneFold Hits:

1ita_ CYTOKINE
Journal Title: STRUCTURE AND FUNCTION OF INTERLEUKIN-1, BASED ON CRYSTALLOGRAPHIC AND MODELING STUDIES
score: **163.1**, score type: **br**, hit name: **1ita_**

1volA COMPLEX(TRANSCRIPTION FACTOR/REGN/DNA)
Journal Title: CRYSTAL STRUCTURE OF A TFIIB-TBP-TATA-ELEMENT TERNARY COMPLEX
score: **134.1**, score type: **br**, hit name: **1volA**

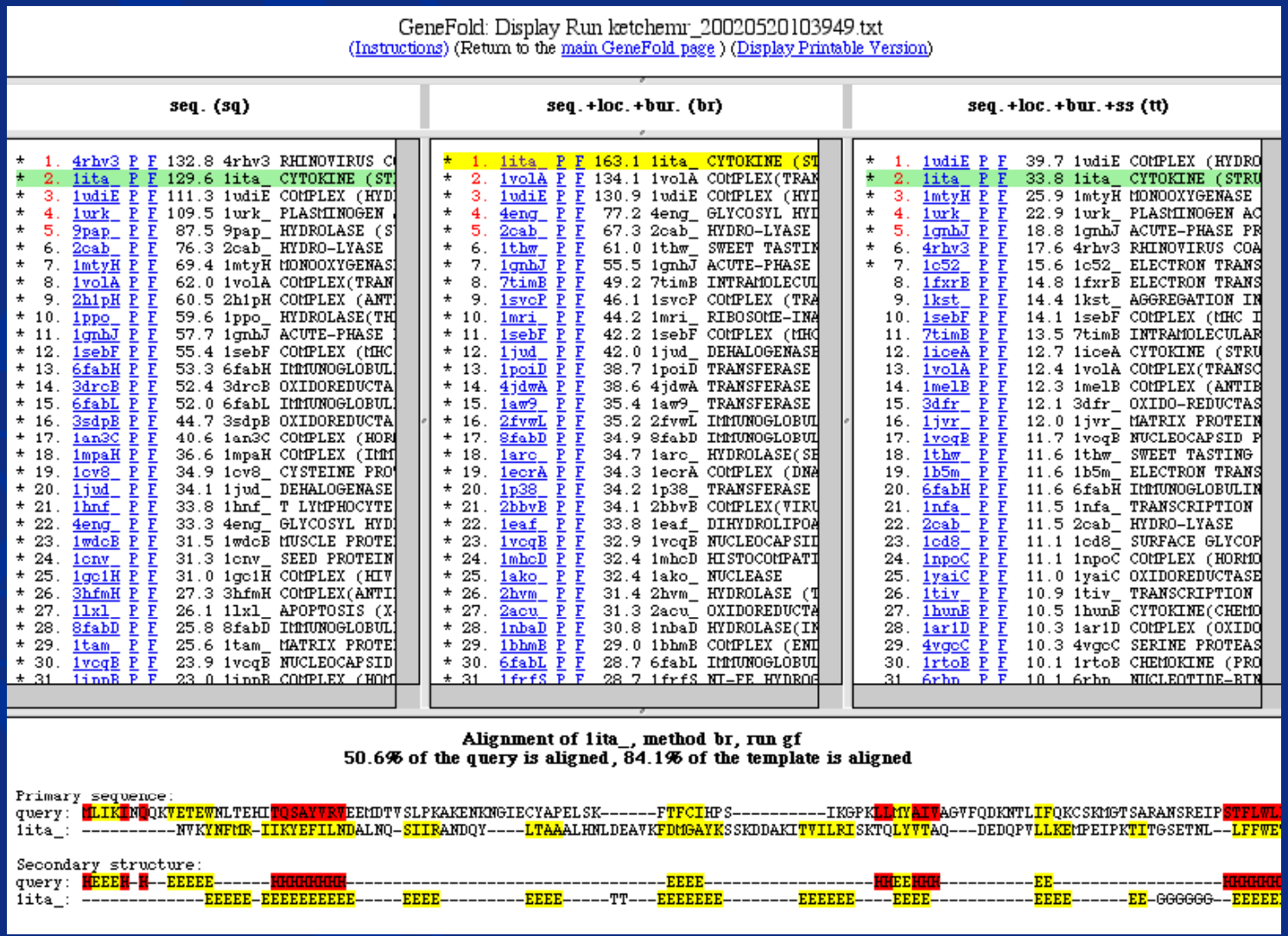
4rhv3 RHINOVIRUS COAT PROTEIN
Journal Title: THE USE OF MOLECULAR-*REPLACEMENT PHASES FOR THE REFINEMENT OF THE HUMAN RHINOVIRUS 14 STRUCTURE
score: **132.8**, score type: **sq**, hit name: **4rhv3**

1udiE COMPLEX (HYDROLASE/INHIBITOR)
Journal Title: NUCLEOTIDE MIMICRY IN THE CRYSTAL STRUCTURE OF THE URACIL-DNA GLYCOSYLASE - URACIL GLYCOSYLASE INHIBITOR PROTEIN COMPLEX
score: **130.9**, score type: **br**, hit name: **1udiE**

1ita_ CYTOKINE
Journal Title: STRUCTURE AND FUNCTION OF INTERLEUKIN-1, BASED ON CRYSTALLOGRAPHIC AND MODELING STUDIES
score: **129.6**, score type: **sq**, hit name: **1ita_**

Mining the Genome

Verify by viewing full GeneFold run



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

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

immunex®

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

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

How SVM's work

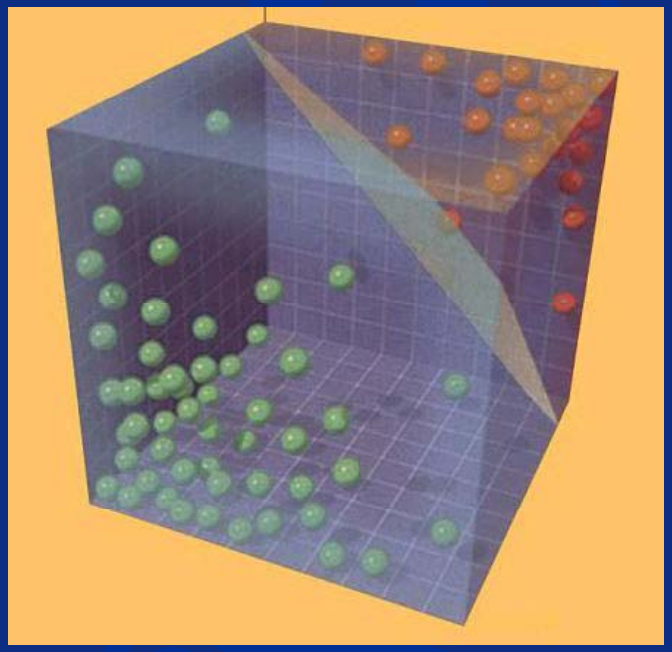


Figure from cover of book: Introduction to Support Vector Machines. Christianni and Shaw-Taylor

SVM work by Paul McDonagh, Immunex Corporation

- 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

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

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

immunex®

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