

COMPUTATIONAL STRUCTURAL BIOLOGY

STRUCTURE, SIMULATION, FUNCTION & PREDICTION

Lecture 7

Michael Levitt
Structural Biology, Stanford

<http://csb.stanford.edu/class>

BIOINFORMATICS II

Structure Comparison.

Structure and Sequence.

Structural Genomics.

Expression Patterns.

Discovering Drugs.

Diagnosing Disease.

Structure Comparison Concept 7.1

STRUCTURE COMPARISON

Structure Superposition.

Gapped Superposition.

Structal.

STRUCTURE SUPERPOSITION

(1) Get an equivalence of points in structure A with structure B.
For example, if these are two forms of the same proteins,
then atom i of A is equivalenced to atom i of B.

(2) Superimpose the coordinates of of B on those of A using.

$$r'_B = T r_B + t,$$

where T is a rotation matrix and t is a translation vector.

(3) Determine T and t to minimize $\sum (r_A - r'_B)^2$.

Dozens of papers have been written on this since 1926.

Diamond, Acta Cryst.
A21, 253. (1966).

McLachlan, Acta Cryst A,
28, 656. (1972)

Kabsch, Acta Cryst.
A32, 922. (1976).

SIMPLE STRUCTURAL ALIGNMENT



DISTANCE AND SIMILARITY

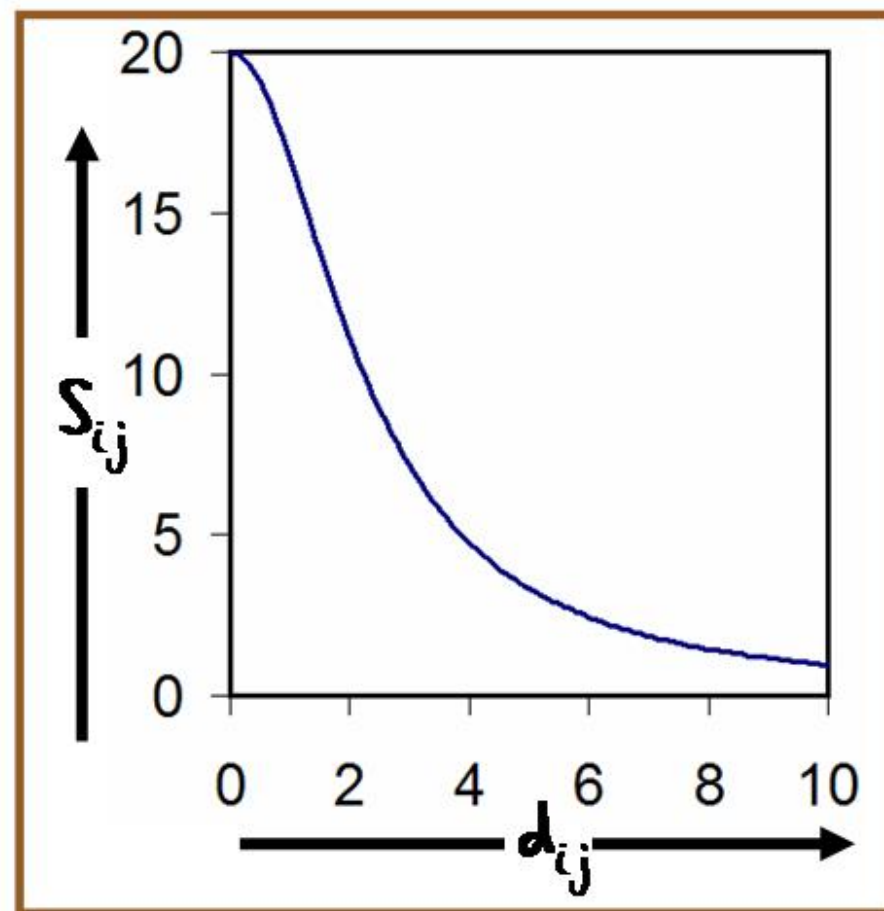
Convert distance between CA atoms i and j to similarity of i and j .

$$S_{ij} = 20 / (1 + d_{ij}^2 / 5)$$

If $d_{ij} = 0$, $S_{ij} = 20$

If $d_{ij}^2 = 5$, $S_{ij} = 10$

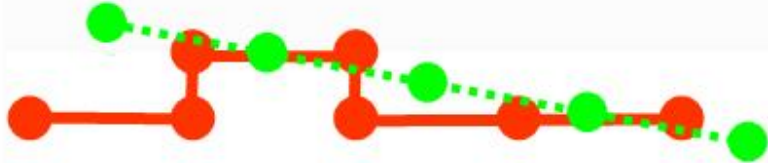
Many possible $S_{ij} = f(d_{ij})$.



STRUCTURAL IN ACTION

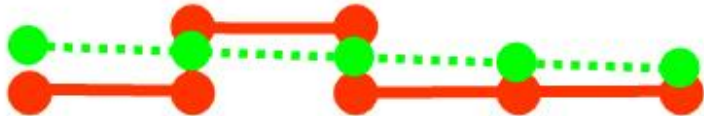
```

- - 1 2 3 4 5    rms 1.96
   | | | | |    brks  0
  1 2 3 4 5 6 7    score 56
    
```



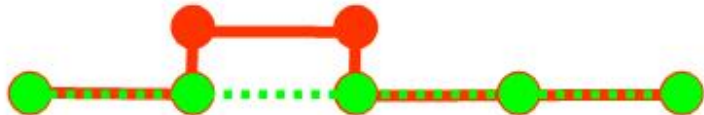
```

1 - 2 - 3 4 5    rms 0.65
| | | | |    brks  2
1 2 3 4 5 6 7    score 57
    
```



```

1 2 - - 3 4 5    rms 0.23
| |   | | |    brks  1
1 2 3 4 5 6 7    score 91
    
```



```

1 2 - - 3 4 5    rms 0.23
| |   | | |    brks  1
1 2 3 4 5 6 7    score 96
    
```



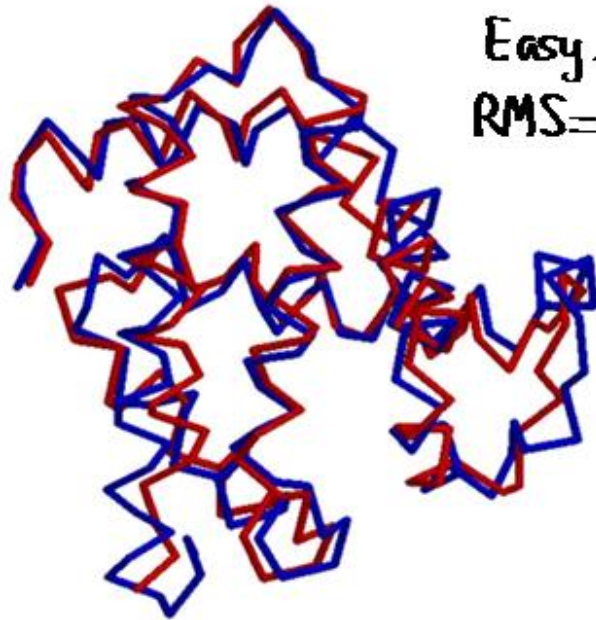
Subbiah et al., Current Biol. 3, 141 (1993).

	1	2	3	4	5	6	7
1	7	5	9	2	1	0	0
2	2	9	12	9	7	2	0
3	1	2	1	10	12	8	2
4	0	1	1	2	2	13	7
5	0	0	0	0	1	2	13

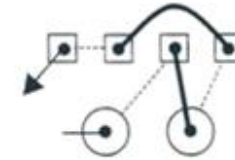
	1	2	3	4	5	6	7
1	19	4	4	1	1	0	0
2	4	16	16	4	4	1	0
3	1	4	4	14	18	4	1
4	0	1	1	4	4	19	4
5	0	0	0	1	1	4	19

	1	2	3	4	5	6	7
1	19	4	3	1	1	0	0
2	4	16	12	4	4	1	0
3	1	4	4	11	19	4	1
4	0	1	1	4	4	19	4
5	0	0	0	1	1	4	19

STRUCTURAL ALIGNMENT

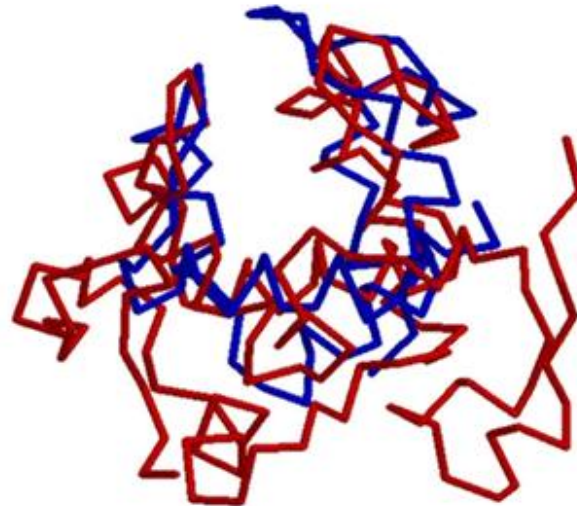


Easy. Two globins.
RMS=1.6Å for 135

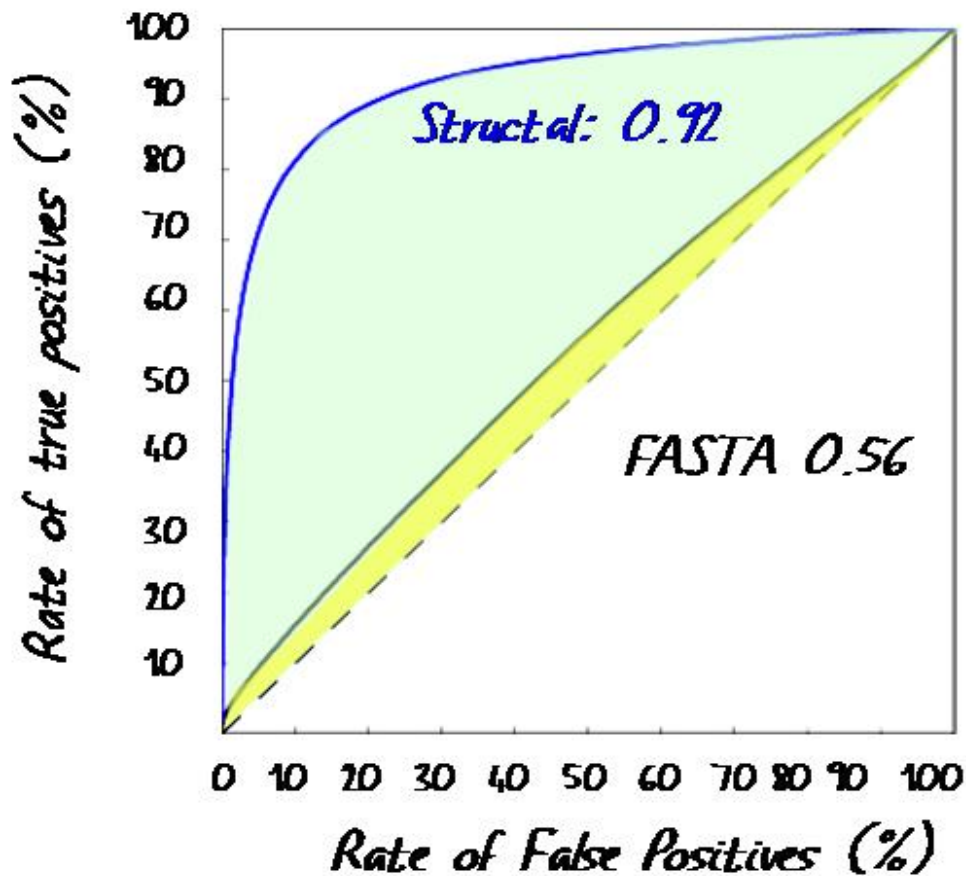


Very Hard.
Same topology
for sub-
domain
(A. Murzin).

Harder.
RMS= 4.3Å for
67.



PROTEIN STRUCTURE SIMILARITY MEASURES



Structural alignment recognizes much more of CATH than sequence alignment does.

Structure and Sequence

Concept 7.2

COMPARING SEQUENCE & STRUCTURE COMPARISON

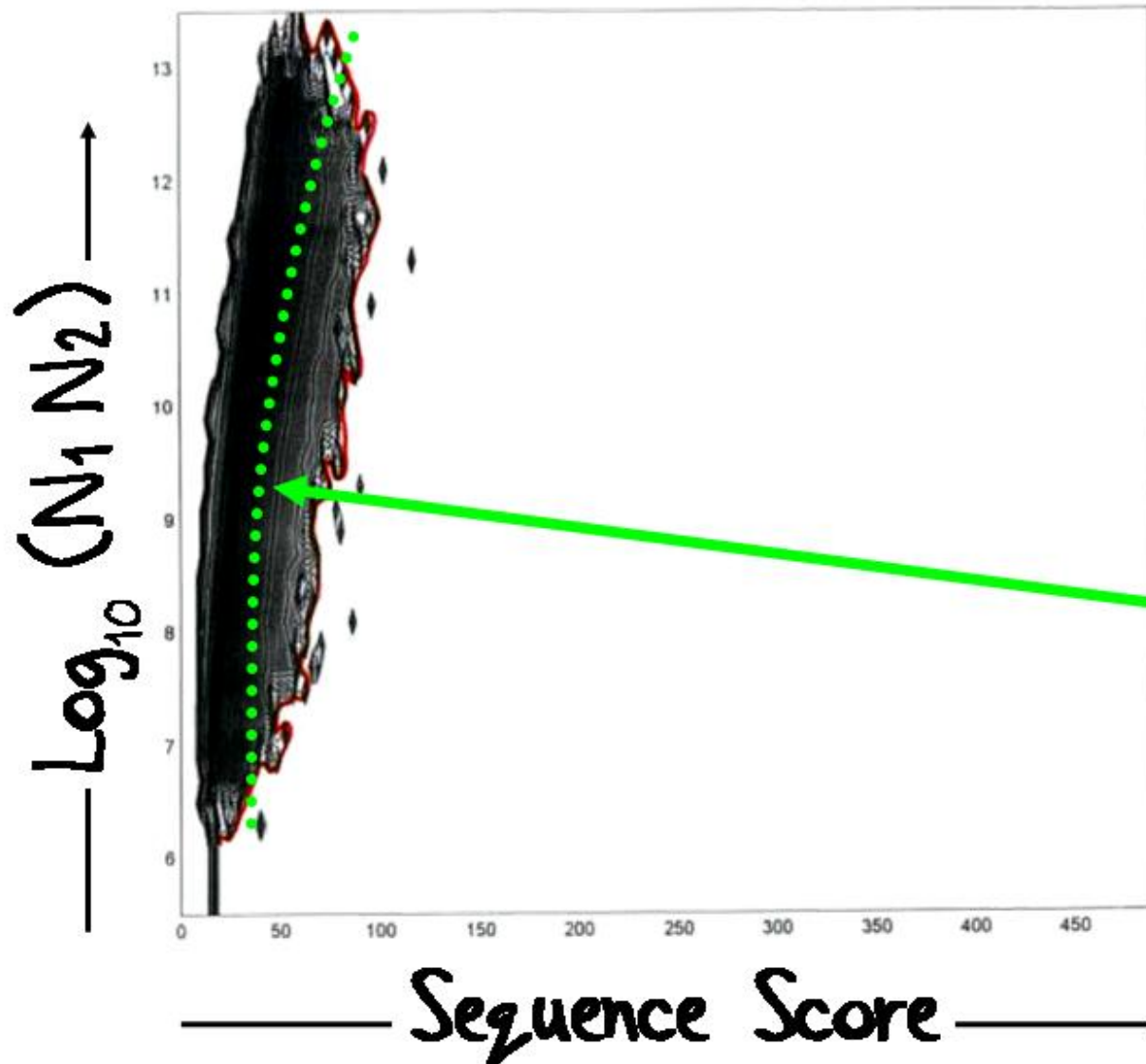
Statistical Significance.

E-values.

Similarity measures.

SEQUENCE SCORE DISTRIBUTION

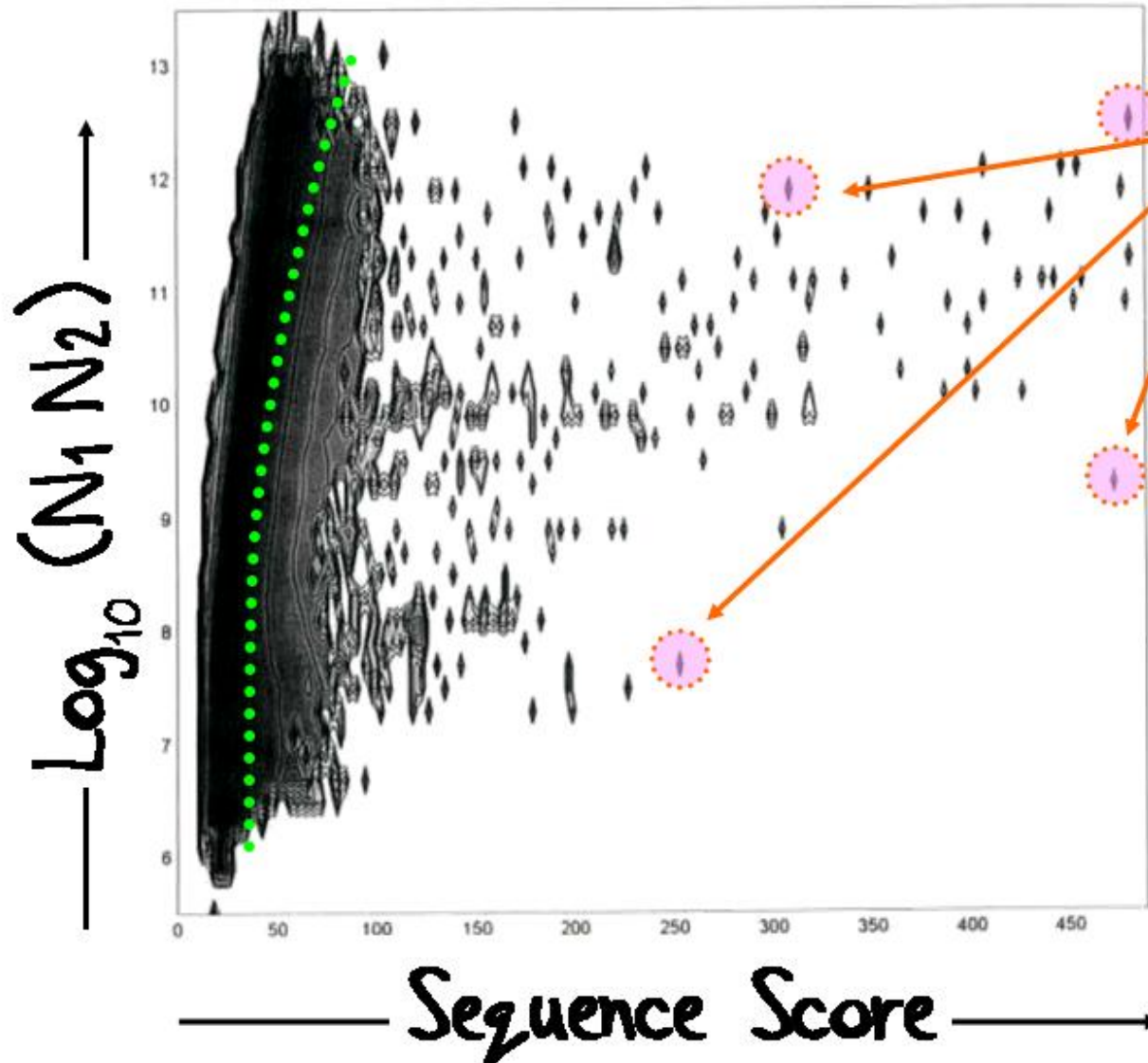
Contour number of occurrences with particular (Score, N) value.



- Include only those pairs of sequences for proteins in different SCOP class.
- None of these pairs should be significantly similar.
- As the protein length increases, the sequence score for a non-significant match increases as $\log n$. (n is sequence length)

SEQUENCE SCORE DISTRIBUTION

Contour number of occurrences with particular (Score, N) value.



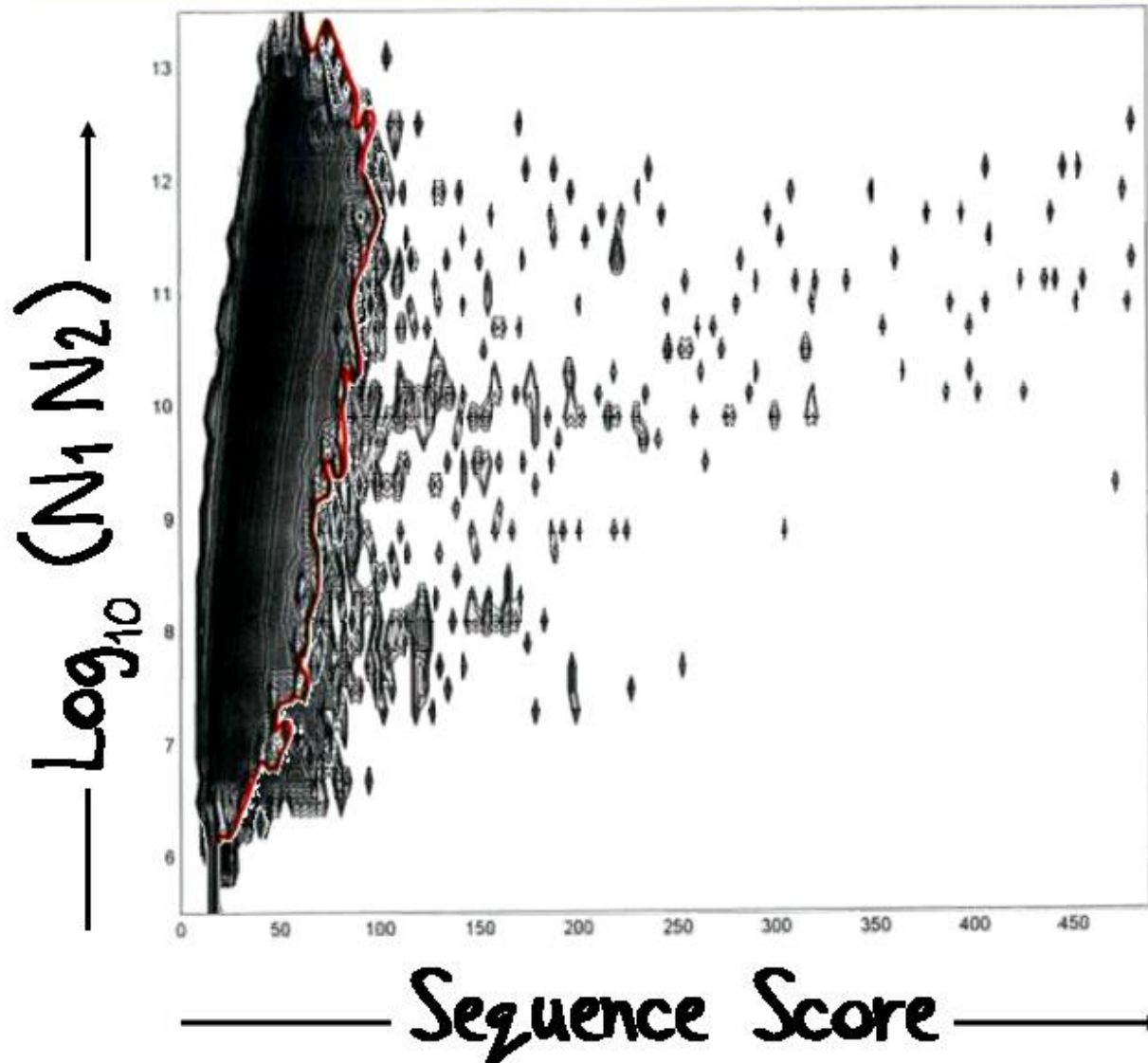
These "islands" are the true positives.

- Include all pairs of sequences.

Levitt & Gerstein, PNAS, 95, 5913 (1998).

SEQUENCE SCORE DISTRIBUTION

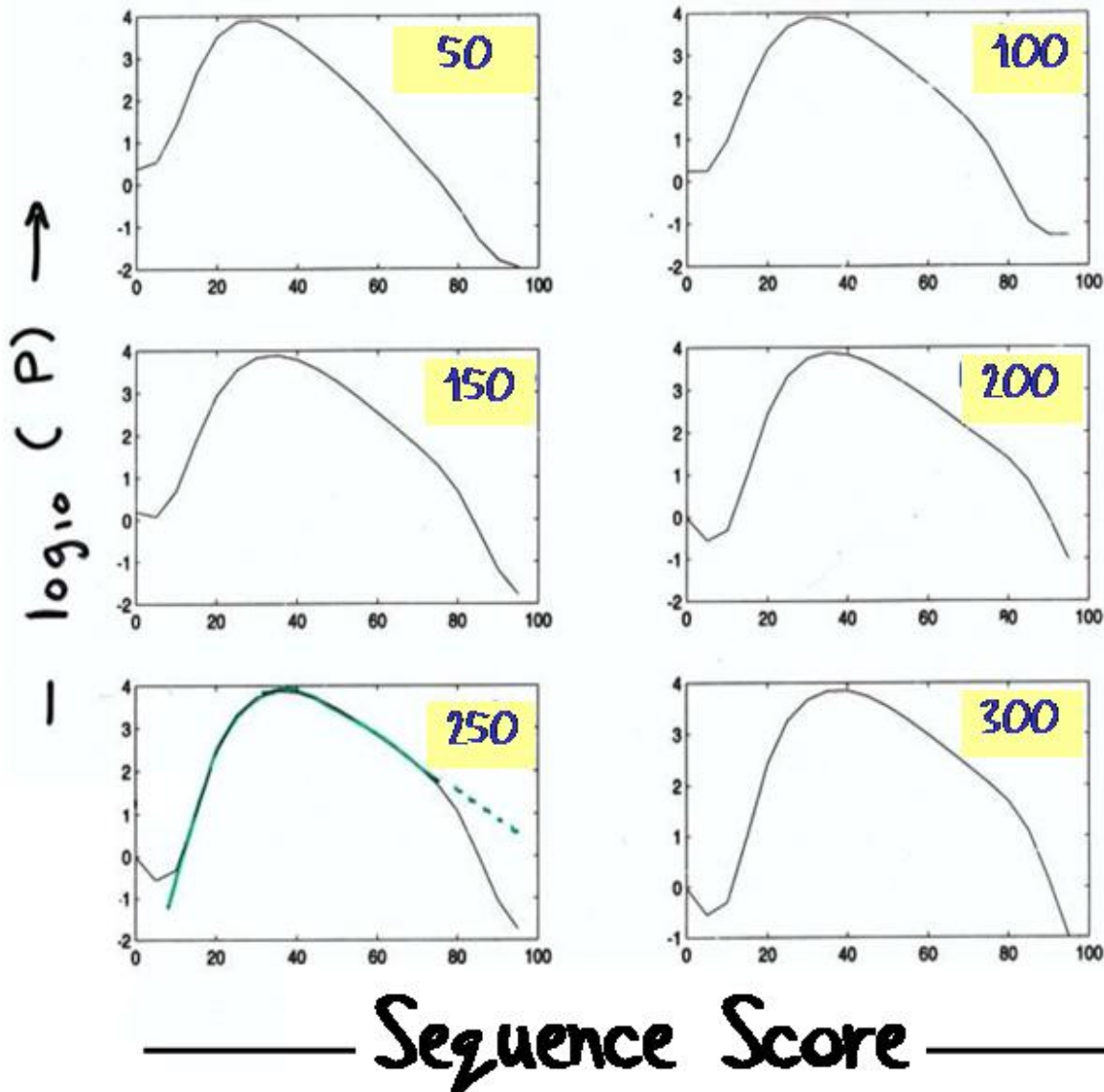
Contour number of occurrences with particular (Score, N) value.



- Show additional pairs in same class.
- Just pairs between SCOP classes.

SEQUENCE SCORE IS EXTREME VALUE DISTRIBUTION

N is sequence length.



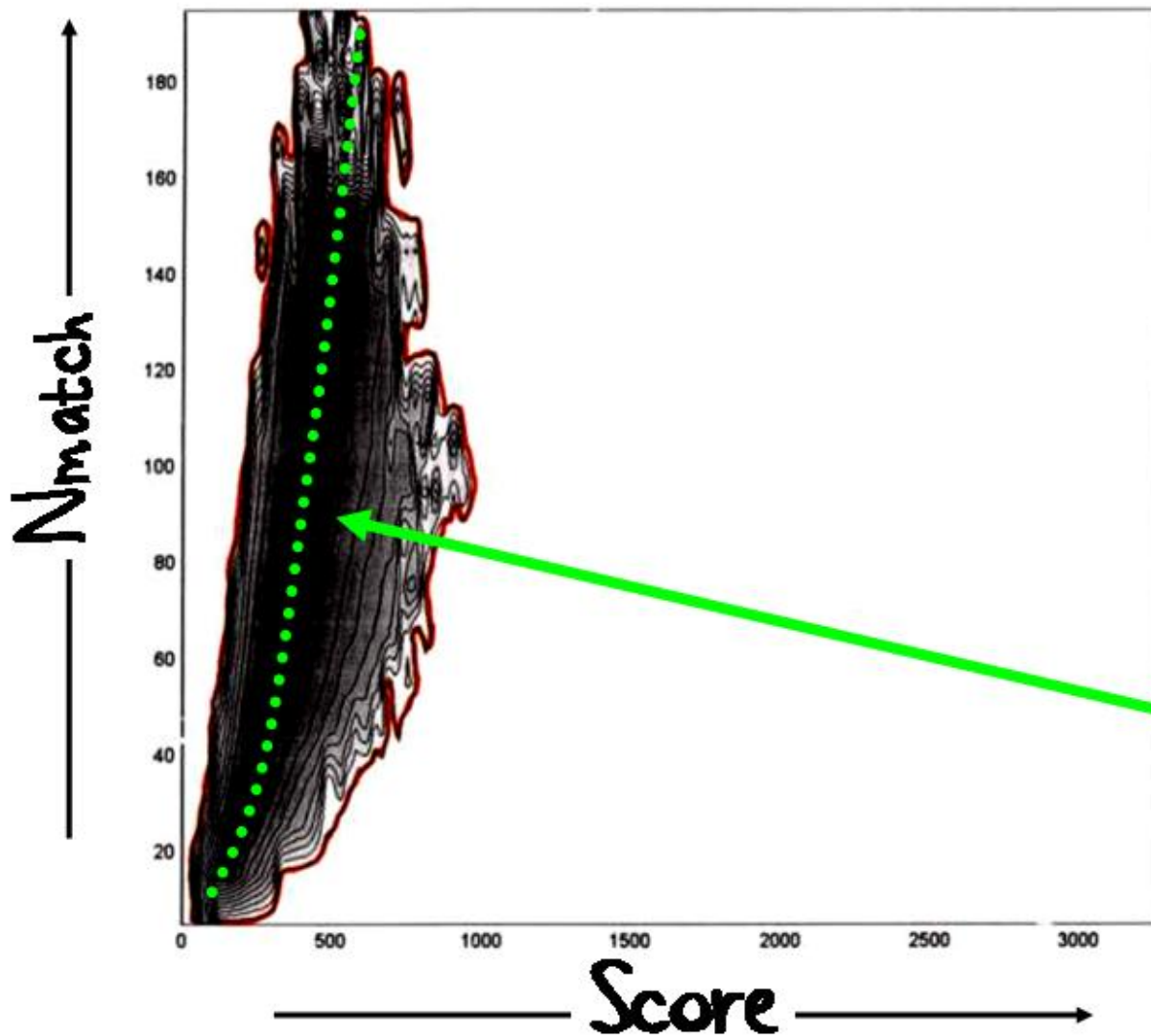
- For the extreme value distribution,
 $P(Z) = \exp(-Z - \exp(-Z))$
 $\log_e P(Z) = -Z - \exp(-Z)$

- For $Z > 0$:
 $\log_e P(Z) = -Z$

- For $Z < 0$:
 $\log_e P(Z) = -\exp(-Z)$.
 where $Z = (\text{score} - \text{mean}) / \text{SD}$.

STRUCTURE SCORE DISTRIBUTION

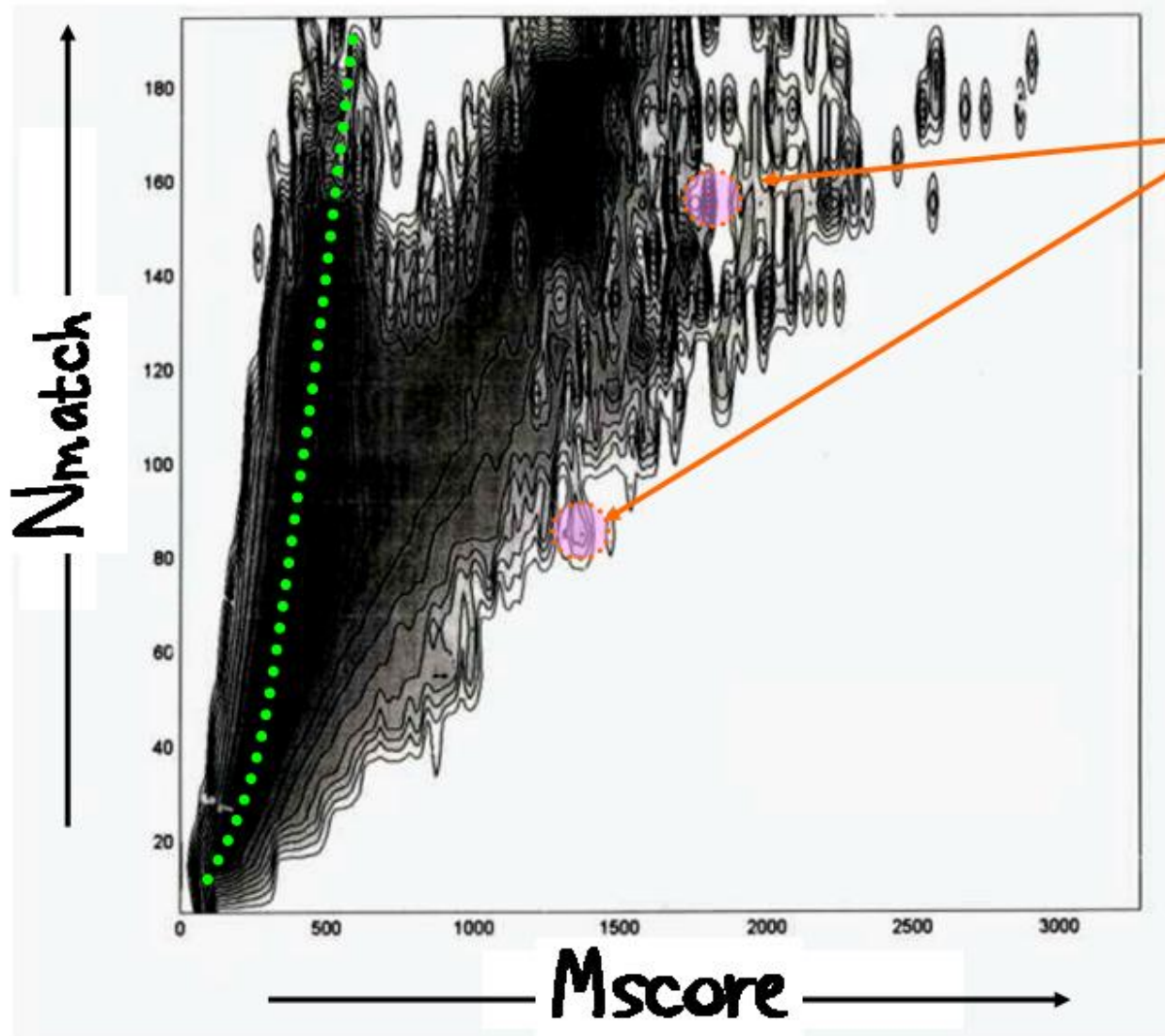
Contour number of occurrences with particular (Score, N) value.



- Include only those pairs of sequences for proteins in different SCOP class.
- None of these pairs should be significantly similar
- As the protein length increases, the Score for a non-significant match increases with N, the match length.

STRUCTURE SCORE DISTRIBUTION

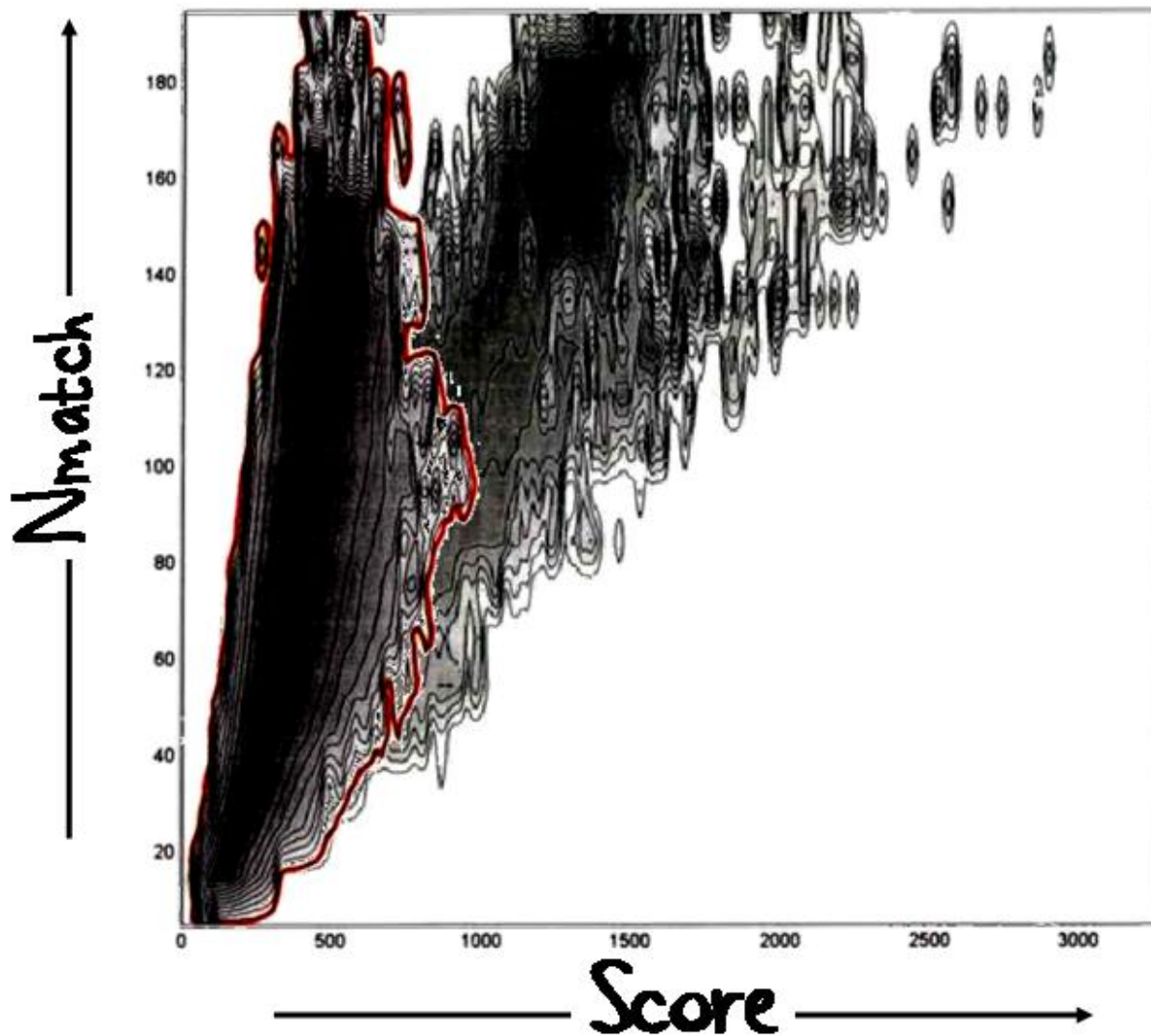
Contour number of occurrences with particular (Score,N) value.



- These "islands" are the true positives.
- Include all pairs of structures for proteins in SCOP.
- The distribution is not as clean as for sequences.

STRUCTURE SCORE DISTRIBUTION

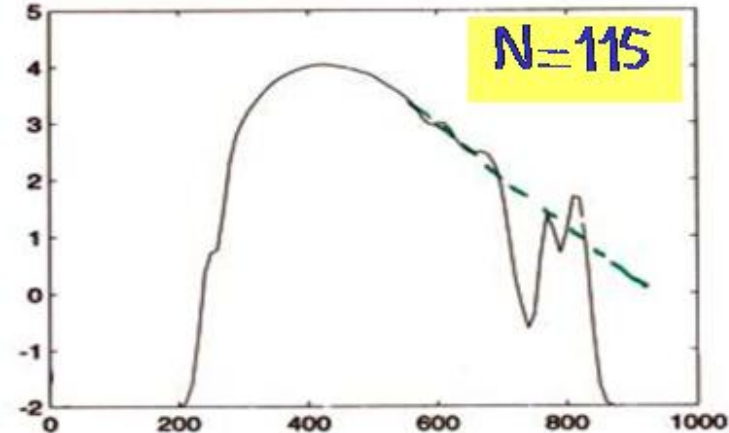
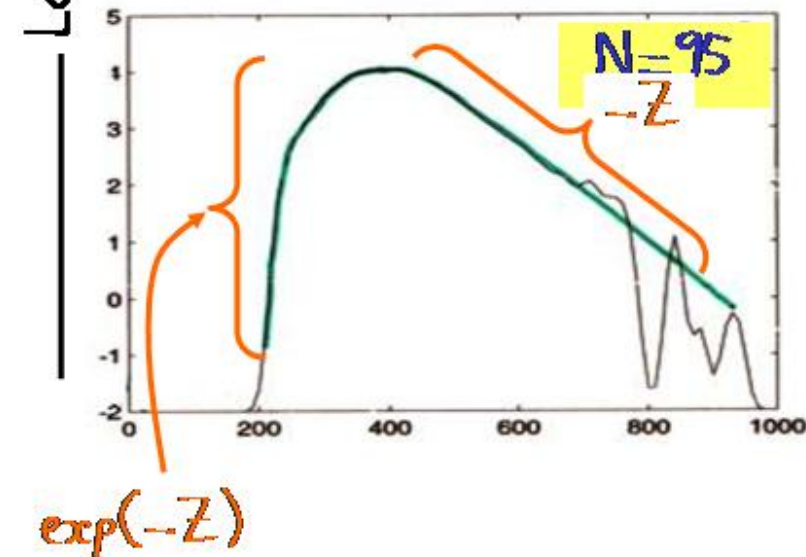
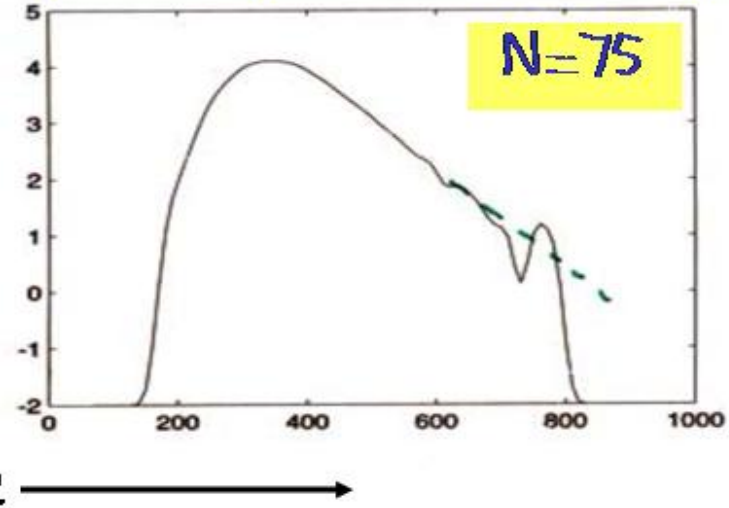
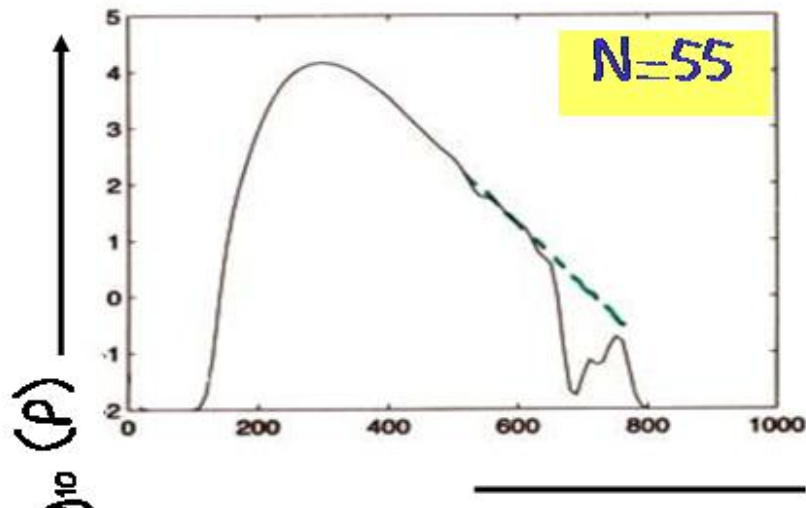
Contour number of occurrences with particular (Score,N) value.



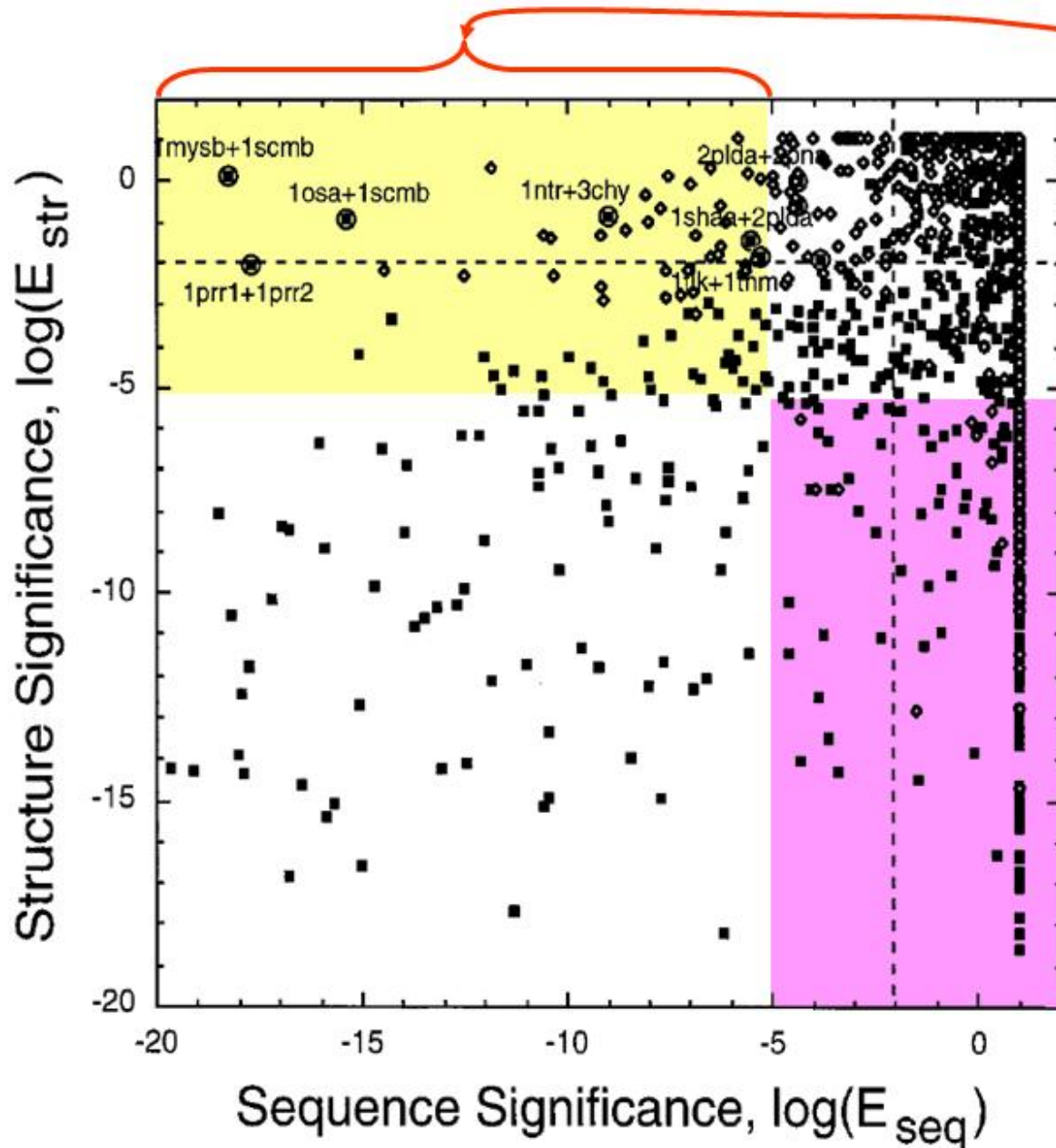
- Show additional pairs in same class.
- Just pairs between SCOP classes.

STRUCTURE SCORE FOLLOWS EXTREME VALUE

N is length of matched region.



STATISTICAL SIGNIFICANCE



• Few pairs have no significant structure match yet have a very significant sequence match.

• Many pairs have no significant sequence match yet have a very significant structure match.

SIMILARITY MEASURES

- The similarity measures that work best for both sequence and structure comparison are a sum of scores for each match.
- These are not the normally used measures.
 - For sequence we like to use percent identity (%ID).
 - For structure we like to use root mean square deviation (RMS).
- Neither %ID or RMS obey an Extreme value distribution for random matches.
- Neither %ID or RMS are reliable indicators of a significant match.

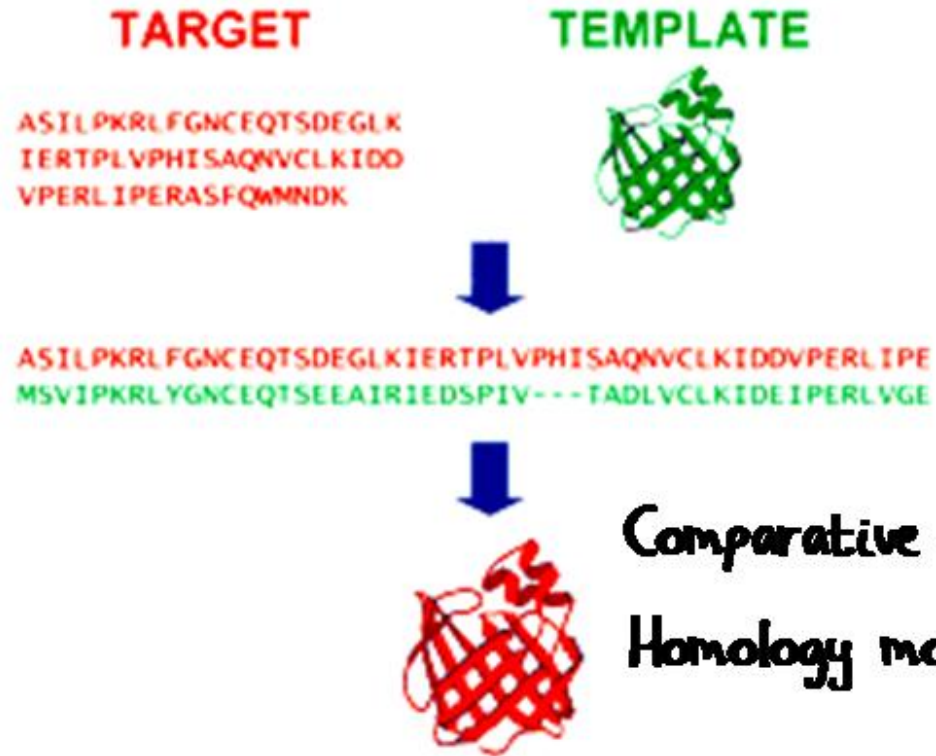
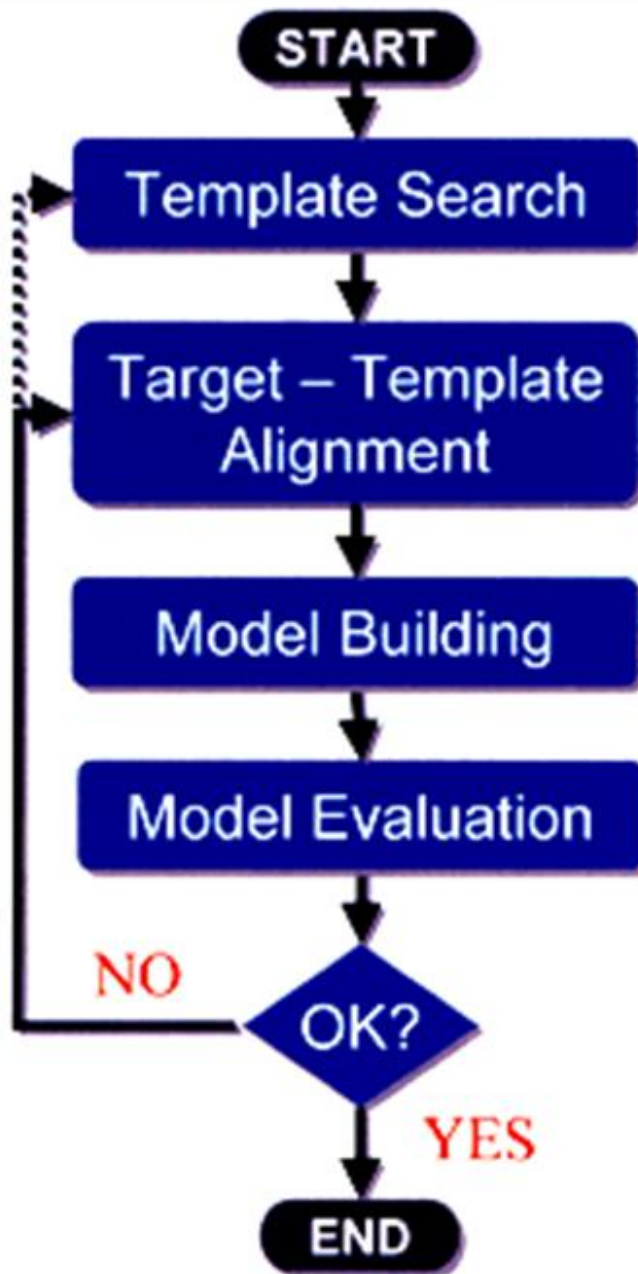
Structural Genomics

Concept 7.3

STRUCTURAL GENOMICS PROJECT

- Aim is to solve structures of all protein sequences.
- This is too much work.
- Solve enough structures so as to be able to model the rest.
- The number that needs to be solved depends on our modeling ability.

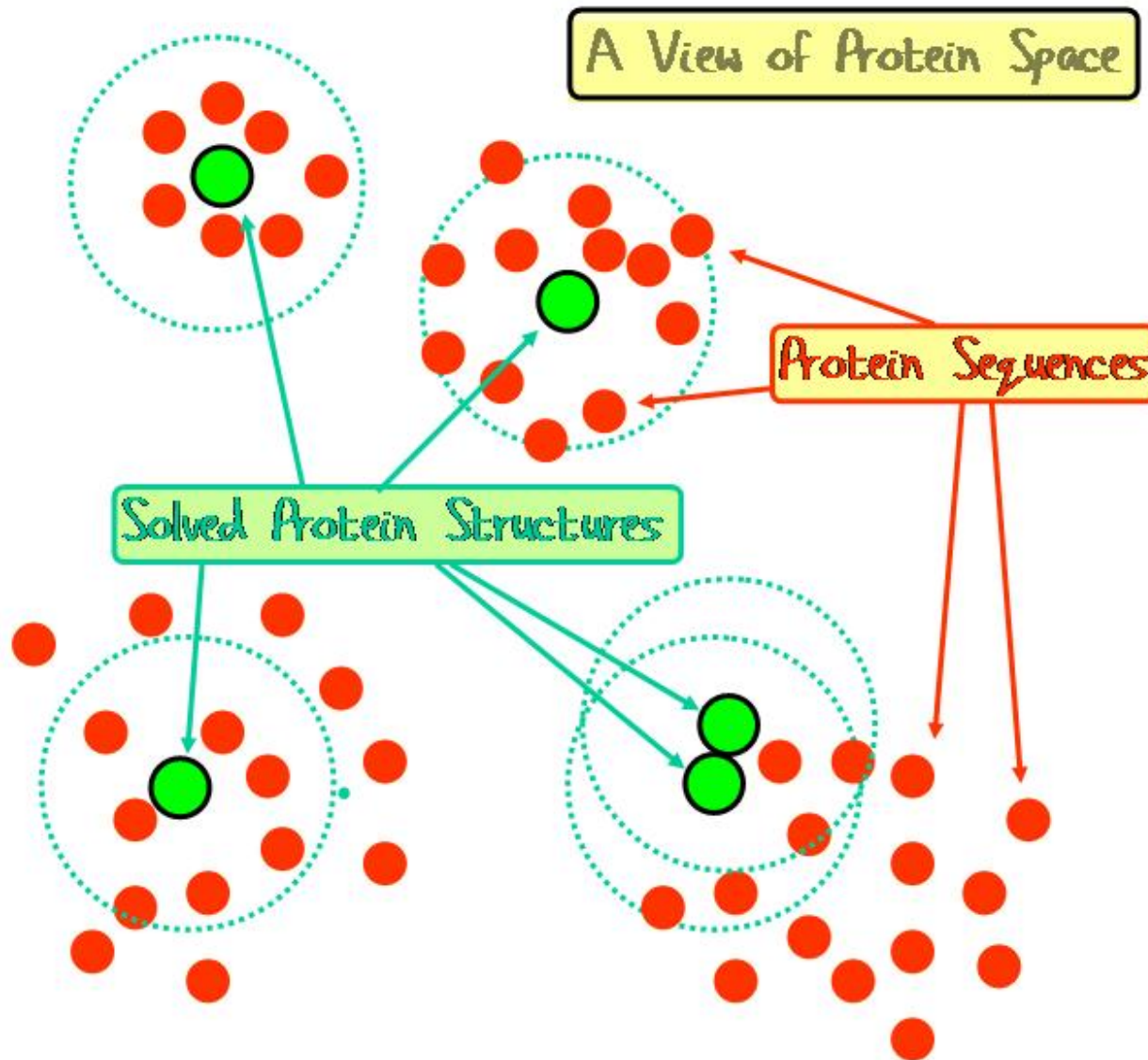
WHAT CAN ONE DO WITH STRUCTURAL GENOMICS



Marti-Rebon, ARBBS, 29, 291 (2000).

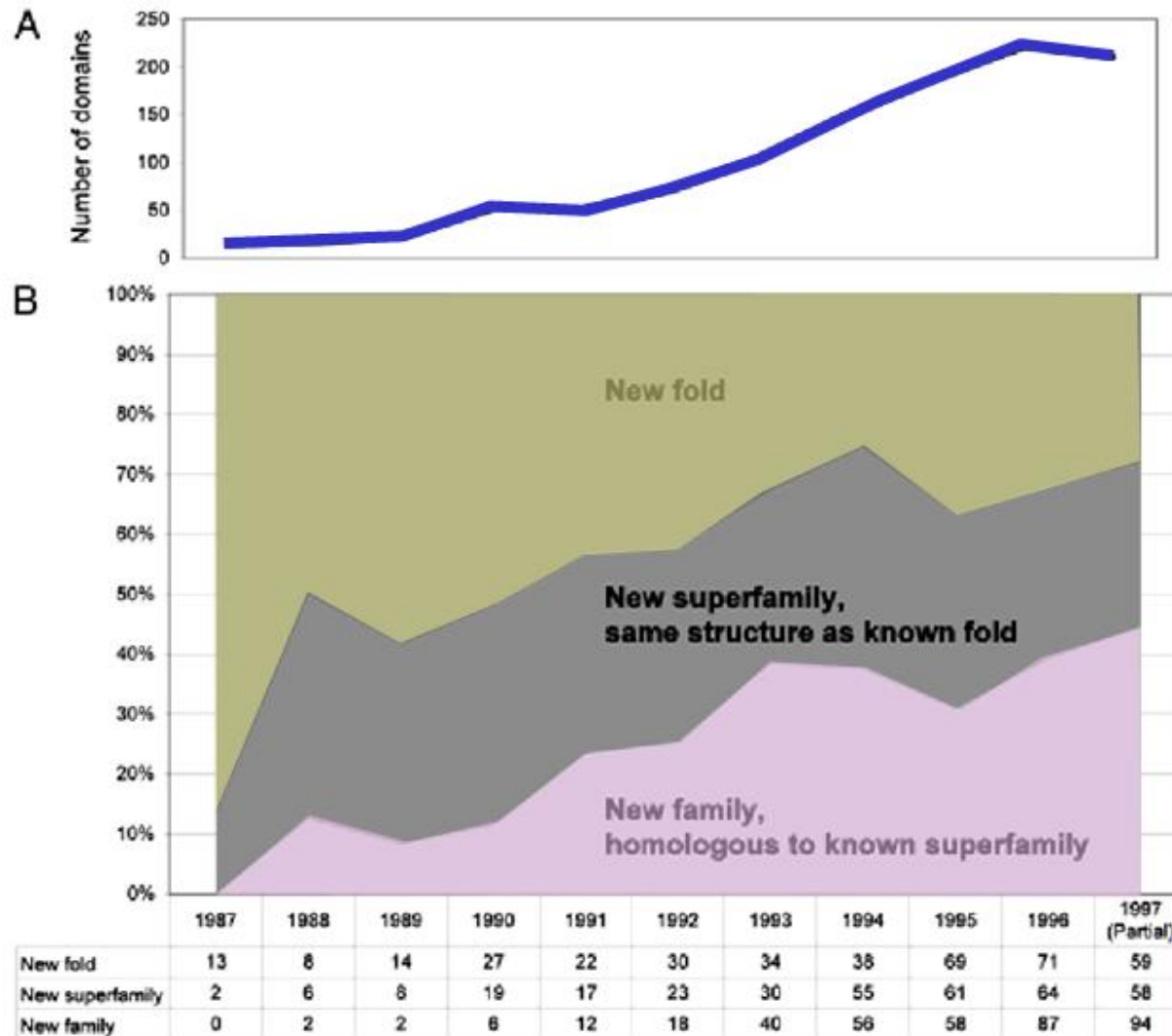
STRUCTURAL GENOMICS & MODELING

A View of Protein Space



- Solve enough structures so as to be able to model the rest.
- The number that needs to be solved depends on our ability to model (the radius of dashed circle).

EXPECTATIONS FROM STRUCTURAL GENOMICS



- This is for sequence that are clearly homologous.
- The fraction of folds that are new was 85% in 1987 and 30% in 1997. Today it is less.

Brenner & Levitt, Protein Engineering, 9, 197 (2000).

JCSG CENTER

JCSG
Joint Center for Structural Genomics

Developing high throughput methods for target selection, cloning, expression, crystallization, X-ray diffraction, and structure determination.

Home | Targets | About | Links | Help | Internal

NEWS & ANNOUNCEMENTS

WEB UPDATES

- ▶ [New Job Postings available](#) - (09-Dec-2003)
- ▶ [Add cod's to job requests](#) - (10-Dec-2003)
- ▶ [Ortholog View](#) - (11-Dec-2003)
- ▶ [Revised "Send to Fold Screening"](#) - (12-Dec-2003)
- ▶ [Search for more than one accession number at a time](#) - (08-Jan-2004)
- ▶ [Target Editor](#) - (16-Jan-2004)
- ▶ [Wild card symbol in target alias filter](#) - (21-Jan-2004)
- ▶ [TSRI Crystal Count](#) - (23-Jan-2004)

LEARN ABOUT US...

- [Organization](#)
- [NIH SG Centers](#)
- [People](#)
- [Careers](#)

PRODUCTION TOOLS...

- [Bioinformatics/Targets](#)
- [Crystallography](#)
- [Structure Validation](#)
- [New Technologies](#)

THINGS TO SEE AND DOWNLOAD...

- [Deposited Structures](#)
- [Target Status](#)
- [Crystal Production](#)
- [Create a Personalized Target List](#)
- [8 New Folds](#)
- [18 Novel Features](#)

JCSG SCOREBOARD

Targets: **6303** Cloned: **2952** Expressed: **2401** Coarse Screen Yield: **660**
Mounted Targets: **286** Data Sets: **134** Structures: **110** Deposited in PDB: **57** New folds: **8**

For information on the Protein Structure Initiative visit [NIHSGS](#) or [StructuralGenomics.org](#)
There are nine sister sites in the SG initiative, for links to the other eight, click [here](#)

[Contact Us](#)

- 8 new folds out of 57 deposited PDB files.

- Look at 6303 targets.

- UCSD, Scripps, Stanford.

BSGC CENTER



BERKELEY STRUCTURAL GENOMICS CENTER

ABOUT BSGC The Berkeley Structural Genomics Center pursues an integrated structural genomics program designed to obtain a near-complete structural complement of two minimal genomes, *Mycoplasma genitalium* and *Mycoplasma pneumoniae*, two related human and animal pathogens.

JOBS

NEWS Working closely with related centers across the country, we strive to present a global view of protein families in nature and to advance new resources for large-scale biological research.

COLLABORATORS Lawrence Berkeley National Laboratory administers the BSGC, in partnership with the University of California, Berkeley, Stanford University and the University of North Carolina, Chapel Hill. We are sponsored by the National Institute for General Medical Sciences of the National Institutes of Health.

WEB RESOURCES

STATUS

CONTACT US

PARTNERS:



Lawrence Berkeley National Laboratory



Stanford University



University of California, Berkeley



University of North Carolina, Chapel Hill

388 targets.

38 in PDB.

New Folds ?

Progress Summary



This page shows how many targets have made it to each experimental stage. It only counts the 'best' form of each target, even if the target has been cloned and expressed multiple times. I.e. if an expressed target is found to be insoluble, and later another expression system makes a soluble form of the target, the target only appears once as 'expressed' and 'soluble'.

Data last updated 2004-01-29 17:44:23 PST8PDT

Experimental Stage	Number of Targets/Structures
Selected	388
Cloned	308
Expression tested	289
Solubility tested	262
Solubility tested (Soluble)	251
Solubility tested (Insoluble)	11
Purified	173
Crystallized	72
Diffraction quality crystals	52
Native diffraction data	23
Phasing diffraction data	44
Traceable map	40
Crystal structure	49 structures / 38 targets
NMR characterization	20
HSQC	19
NMR structure	2 structures / 2 targets
In PDB	44 structures / 33 targets
Biochemical function	15
Work stopped	138

• UCB, UNC, LBL
Stanford

MCSG CENTER

Security / Privacy Notice

MCSG

Midwest Center for Structural Genomics

PSI

Structure Gallery

• XML Files • Target List • Progress • Statistics • Log in • Site Search: Go

Consortium
Project
Investigators
Targets
3-D Structures
Related Publications
SG Sites
NIH
MCSG Resources
Job opportunities
Collaborators
Internals

Consortium Members:


- Argonne National Laboratory
- Northwestern University
- Washington University School of Medicine
- University College London
- UT Southwestern Medical Center at Dallas
- University of Toronto
- University of Virginia

Active Targets: 1325
Crystallized: 318
In PDB: 112
New Folds: 15


The webpage is under construction in W.Minor Lab

1325 targets.
112 in PDB.
15 new folds.

NYSG CENTER



**New York Structural Genomics
Research Consortium**



Mission Statement
To develop and use the technology for high-throughput structural and functional studies of proteins.

Participating Research Groups


Albert Einstein College of Medicine	The Rockefeller University
Brookhaven National Laboratory	University of California, San Francisco
Columbia University	Weill Medical College of Cornell University
Structural Genomics, Inc	


Public Target Information


[Public Target Progress Report](#)
[Download Public Target Progress Report in XML Format](#)

Progress Statistics

Selected:	1713;
Cloned:	759;
Expressed:	688;
Soluble:	626;
Purified:	472;
Crystallized:	178;
Diffraction-quality Crystals:	106;
Native diffraction-data:	83;
Phasing diffraction-data:	83;
Crystal Structure:	83;
Deposited in PDB:	73;


More Statistics


Structure Gallery


Model Coverage by
NYSGXRC Structures

[Home](#) [Proposal](#) [Publications](#) [Flowchart](#) [IceDB](#) [Tools](#) [Contact](#)

• Columbia, BNL, Rockefeller, UCSF, SGX Inc.

1713 targets.
73 in PDB.
? new folds.

TBSG CENTER

Mycobacterium tuberculosis Structural Genomics Consortium

- Join the [listserve](#) and keep updated on the announcements from the Consortium

[Click Genome to browse genes](#) [Consortium Member Login](#)

[Community Member Login](#)

Home [News & Research Highlights](#) [Consortium Overview](#) [Strategy](#) [Consortium Members](#) [Sponsors](#)
[Consortium Policies](#) [Targets and Results](#) [Community Pages](#) [TB ORF Info Place](#) [Selected Publications](#) [Job Opportunities](#)
[Links to Related Sites](#) [Assisted PubMed Search](#) [Site Map](#) [Mycobacteriophage Page](#)
[Software](#)

• UCLA,
Los Alamos

Stage	# of Experiments	# of ORFs	ORFs worked on by Facilities
Targeted	4077	1472	555
Cloned	2955	1212	452
Cloned II (Exp Vect)	2482	1122	429
Expressed	1996	858	330
Solubility I	1252	536	101
Purified	808	311	73
Solubility II	569	240	51
In Crystal Trials	546	223	51
Crystallized	218	121	36
Diffracting Xtal	146	85	17
Data Collected	131	78	15
Phased	98	61	4
Model Built	96	61	4
Refined	73	44	3
Deposited	43	29	3

4077 targets.
43 in PDB.
? new folds.

STRUCTURAL GENOMICS OUTPUT

JOINT GENOME	57	8
BERKELEY	38	?
MIDWEST	112	15
NEW YORK	73	?
TUBERCULOSIS	43	?

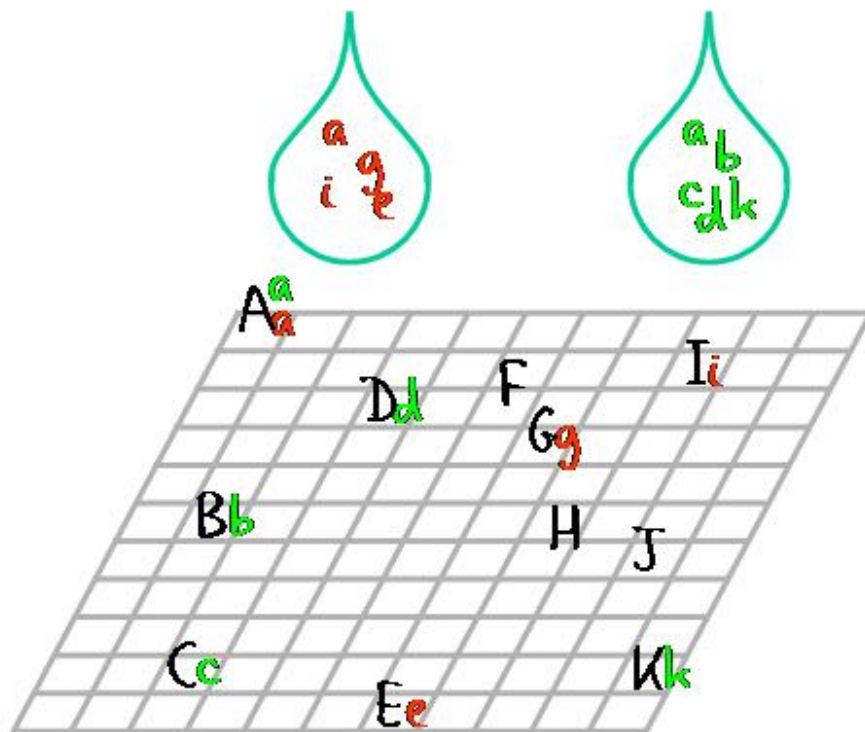
Estimate 15% new so about 50 new folds.

Expression Patterns

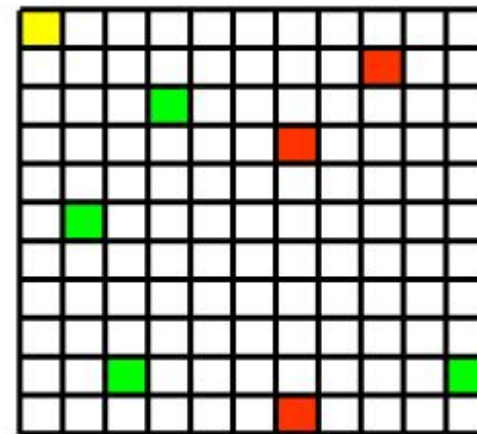
Concept 7.4

MICROARRAY BASICS

- Use position in array to distinguish immobilized molecules (A, B, C, D, E, F, G, H, I, J...).
- Use dye color (red or green) to distinguish source of soluble molecules (a, b, c, d, e, f, g, h, i, j ...).



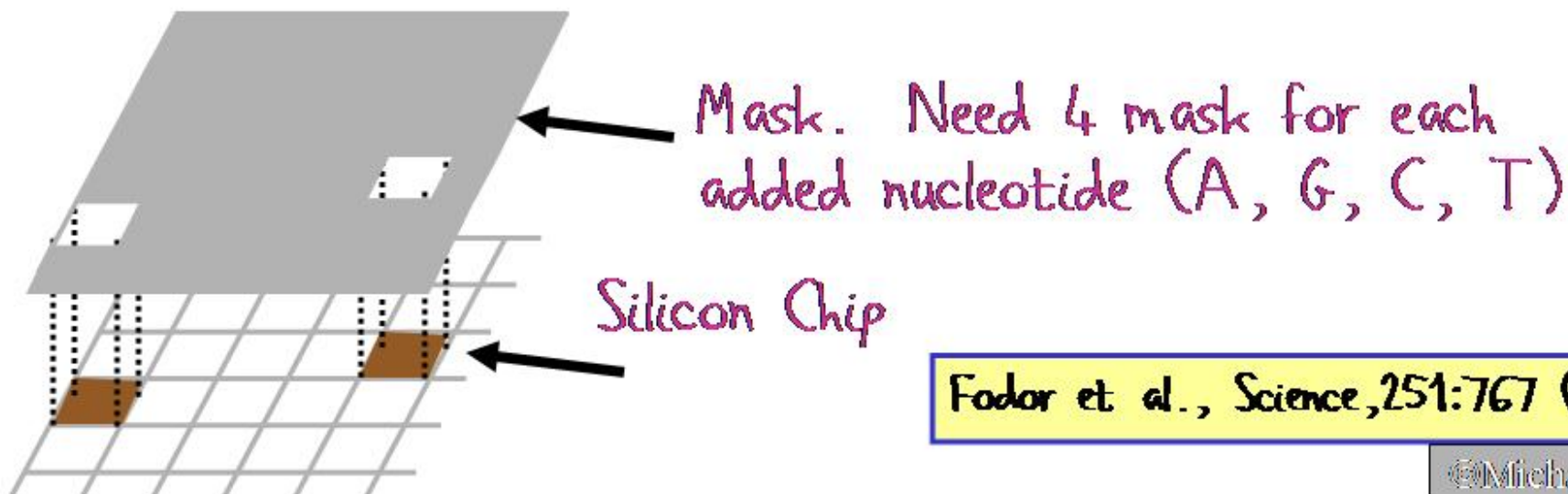
- Assume A only binds a, B only binds b, etc.



- Use fluorescence pattern to analyze.

OLIGO DNA MICROARRAYS

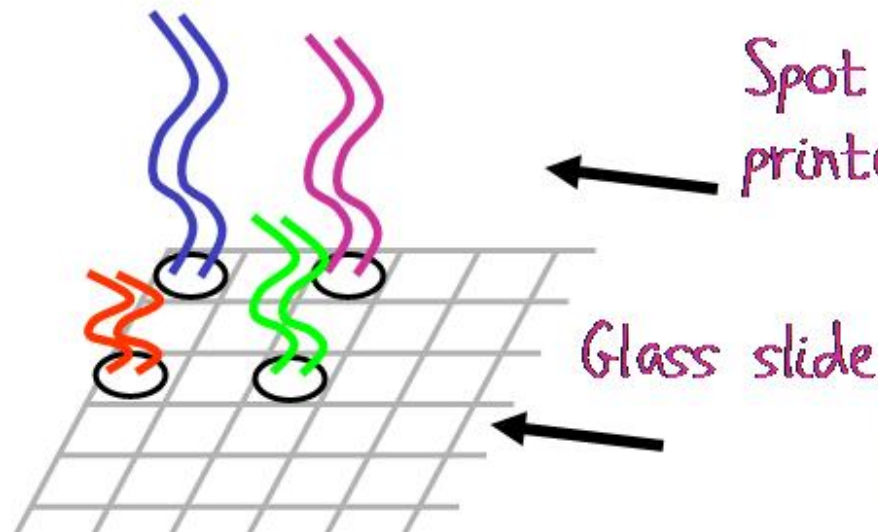
- In DNA arrays, A, B, C, D, etc are short single-stranded DNA oligomers of length ~ 25 .
- Make DNA arrays by photo-lithography.
- Density can be very high (11 μm features).
1,300,000 oligos per chip.
- Need several about 20 oligos per gene.



Fodor et al., Science, 251:767 (1991).

CDNA EXPRESSION ARRAYS

- In cDNA arrays, A, B, C, D, etc are single-stranded cDNA molecules of length ~ 500 or more.
- Make cDNA arrays by spotting or printing.
- Density lower than DNA chip. More uniform
 - 25,000 genes per chip.

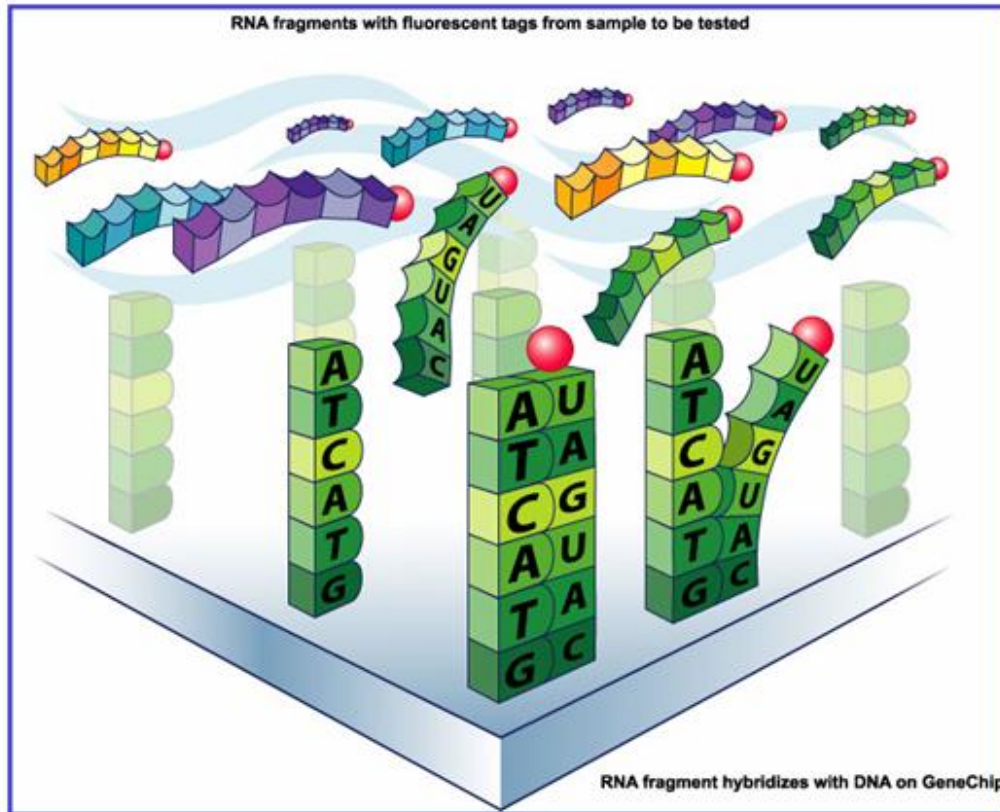


Spot cDNA with robot or ink-jet printer. Make DNA from sequence.

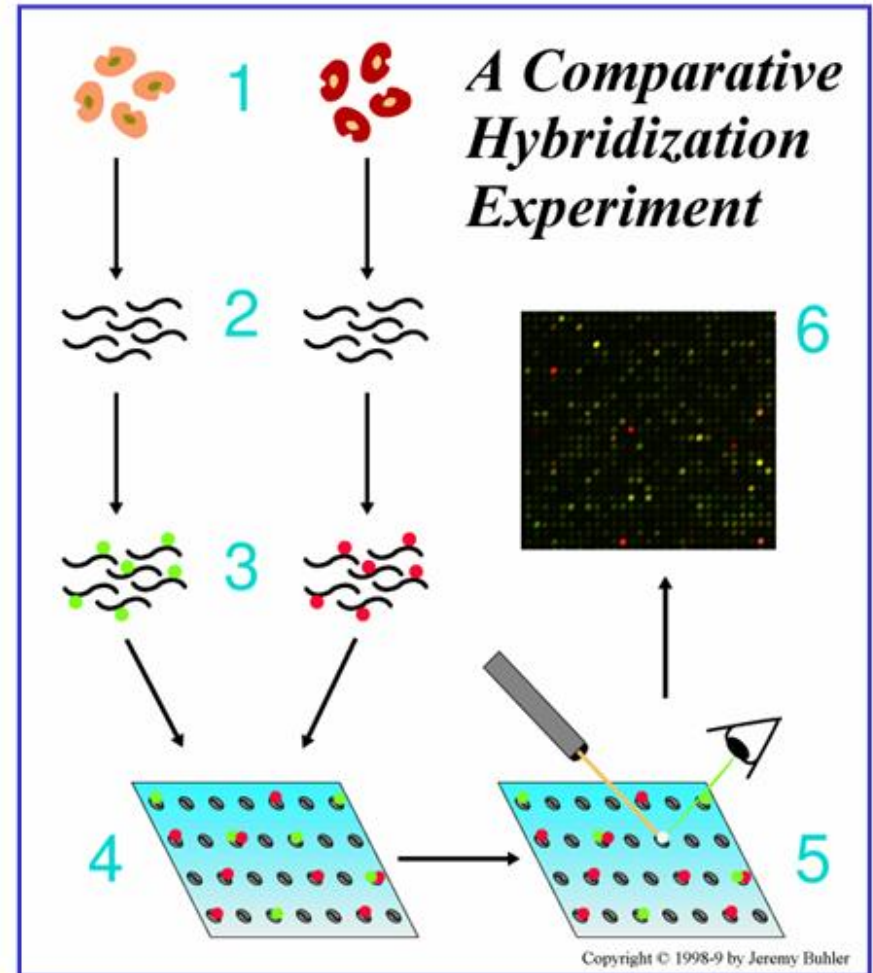
Schena et al. Science, 270: 467 (1995).

MORE EXPRESSION ARRAY CARTOONS

cDNA chip also hybridizes RNA



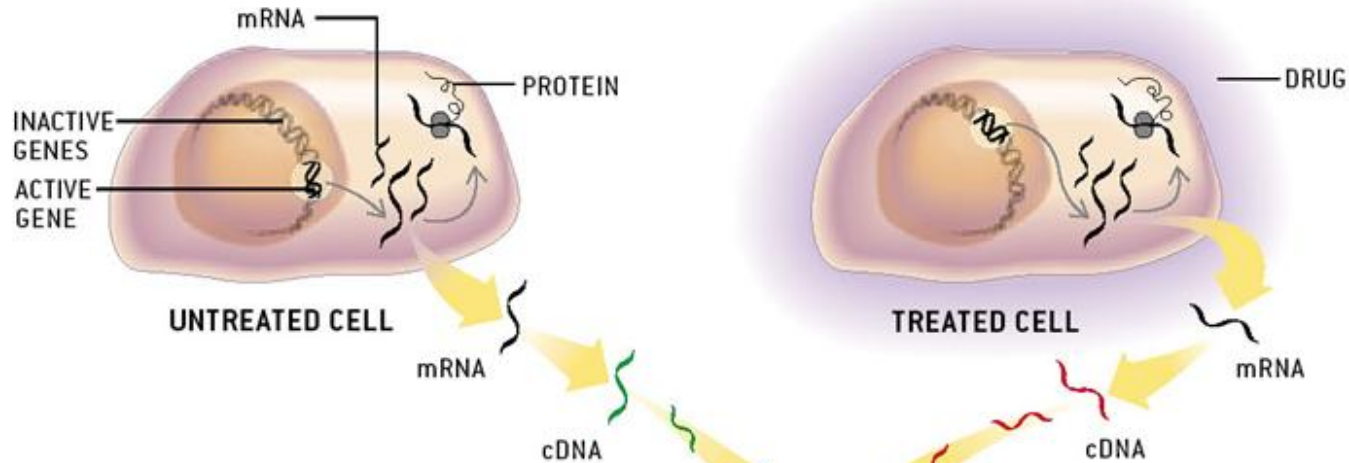
Oligo DNA Chip hybridizes RNA



©Michael Levitt 04

SCIENTIFIC AMERICAN EXPLAINS

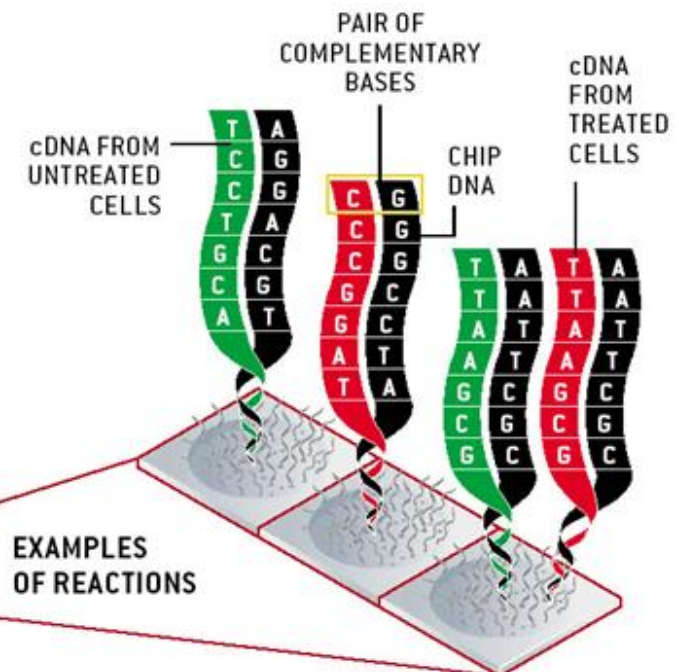
Experimental setup



3 Transcribe the mRNA into more stable complementary DNA (cDNA) and add fluorescent labels—green to cDNAs derived from untreated cells, red to those from treated cells.

4 Apply the labeled cDNAs to the chip. Binding occurs when cDNA from a sample finds its complementary sequence of bases on the chip (*detail at right*). Such binding means that the gene represented by the chip DNA was active, or expressed, in the sample.

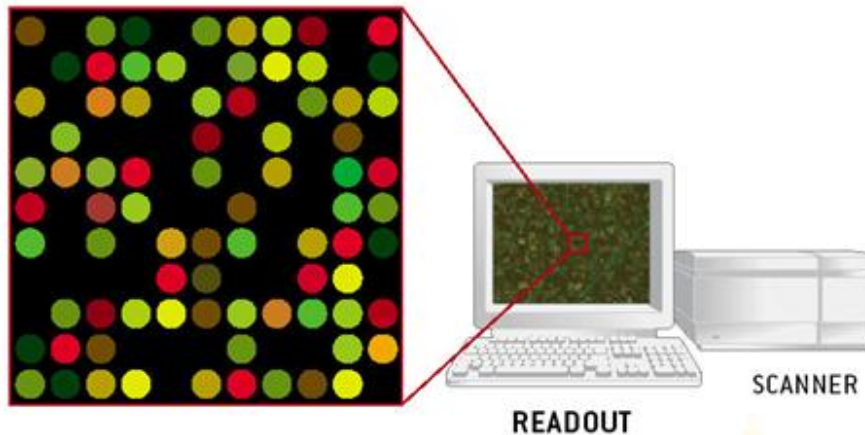
2 Obtain two samples of liver cells; apply the drug to one sample. Then, from each sample, collect molecules of messenger RNA (mRNA)—the mobile copies of genes and the templates for protein synthesis in cells.



SCIENTIFIC AMERICAN EXPLAINS

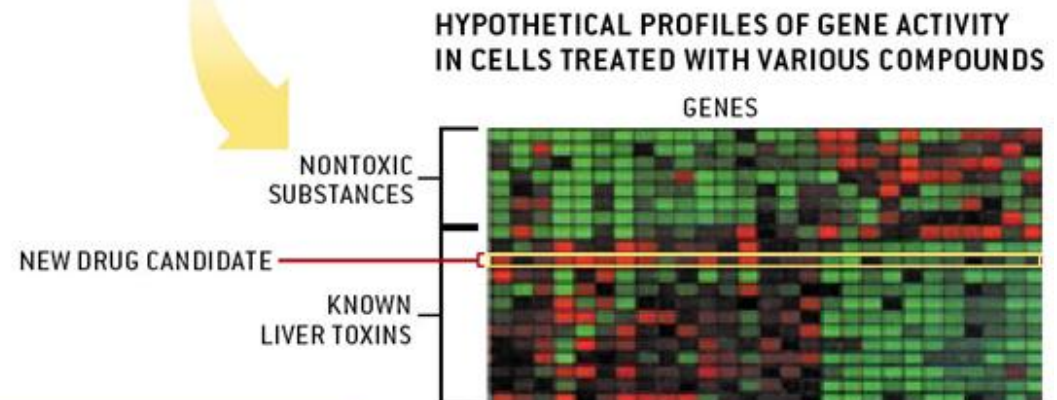
Computational analysis.

- GENE THAT STRONGLY INCREASED ACTIVITY IN TREATED CELLS
- GENE THAT STRONGLY DECREASED ACTIVITY IN TREATED CELLS
- GENE THAT WAS EQUALLY ACTIVE IN TREATED AND UNTREATED CELLS
- GENE THAT WAS INACTIVE IN BOTH GROUPS



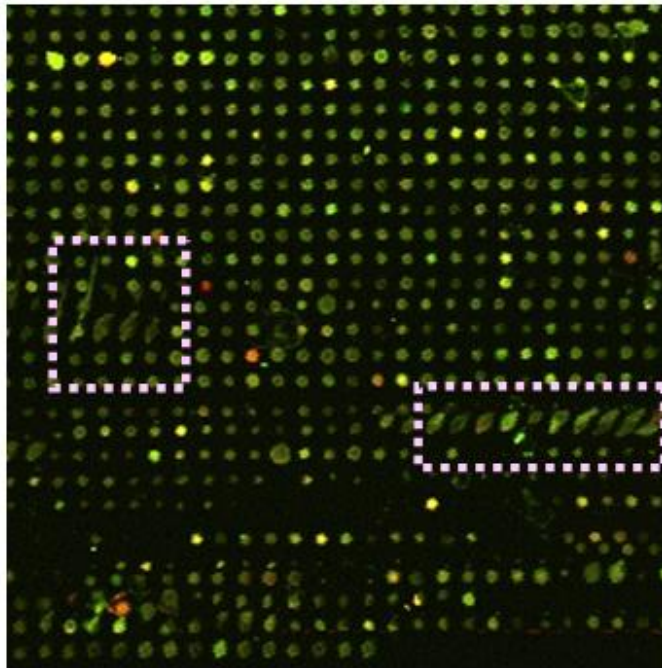
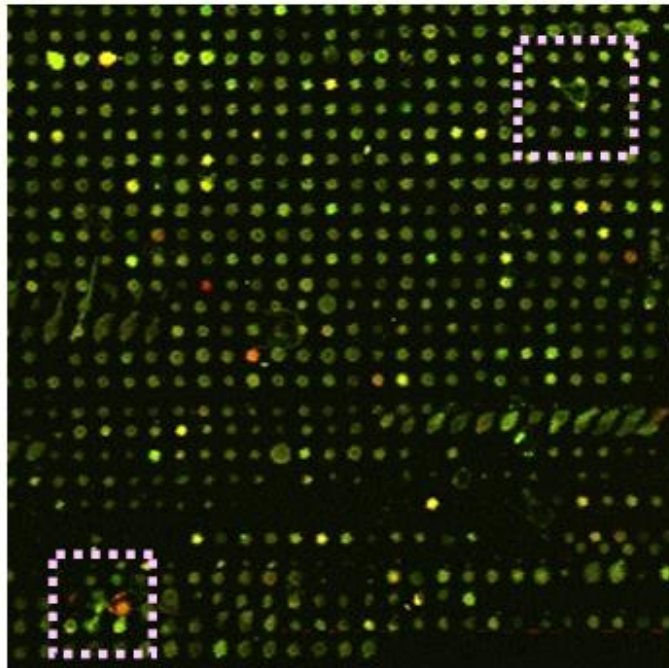
5 Put the chip in a scanner. Have a computer calculate the ratio of red to green at each spot (to quantify any changes in gene activity induced by the drug) and generate a color-coded readout.

6 Determine whether any genes responded strongly to the drug in ways known to promote or reflect liver damage. Or compare the overall expression pattern produced by strong responders with the patterns produced when those genes react to known liver toxins (*right*). Close similarity would indicate that the new candidate was probably toxic as well. In the diagram, each box represents a single gene's response to a compound.



Friend SH, Stoughton RB. The Magic of Microarrays
Sci Am. 286: 44-49 (2002)

CLEANING UP OLIGO CHIPS



Filter bad Spots:

Small spots.

Smudgy spots.

Non-round spots.

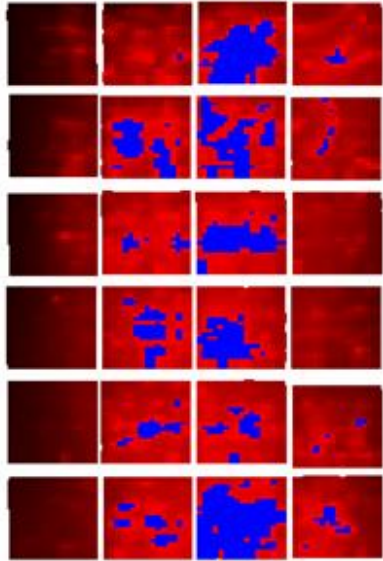
Background correction:

Raw data has large positional dependence.

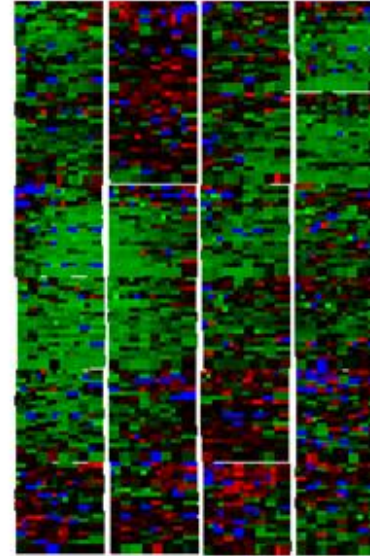
Remove artifactual smudges.

http://bioinfo.mbb.yale.edu/mbb452a/2002/slides_htmls/expression_4.html

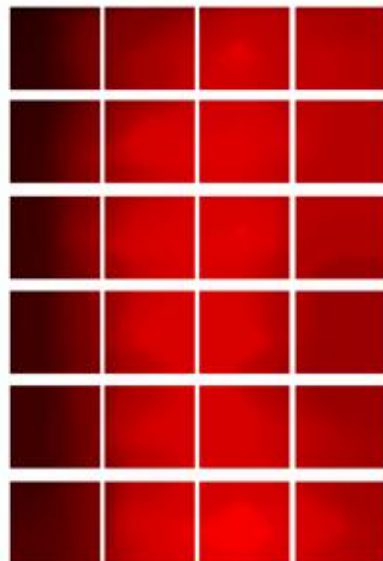
NOISY BACKGROUND



Before

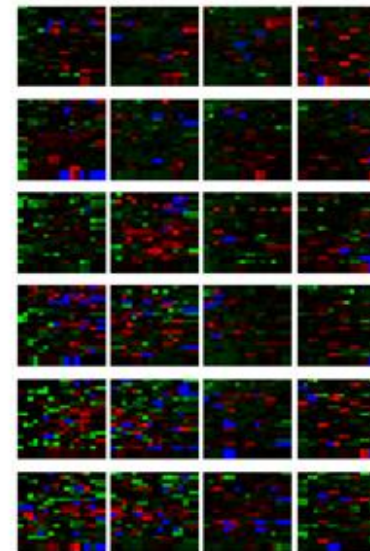


Raw data has large positional dependence



After

Remove
artifactual
smudges

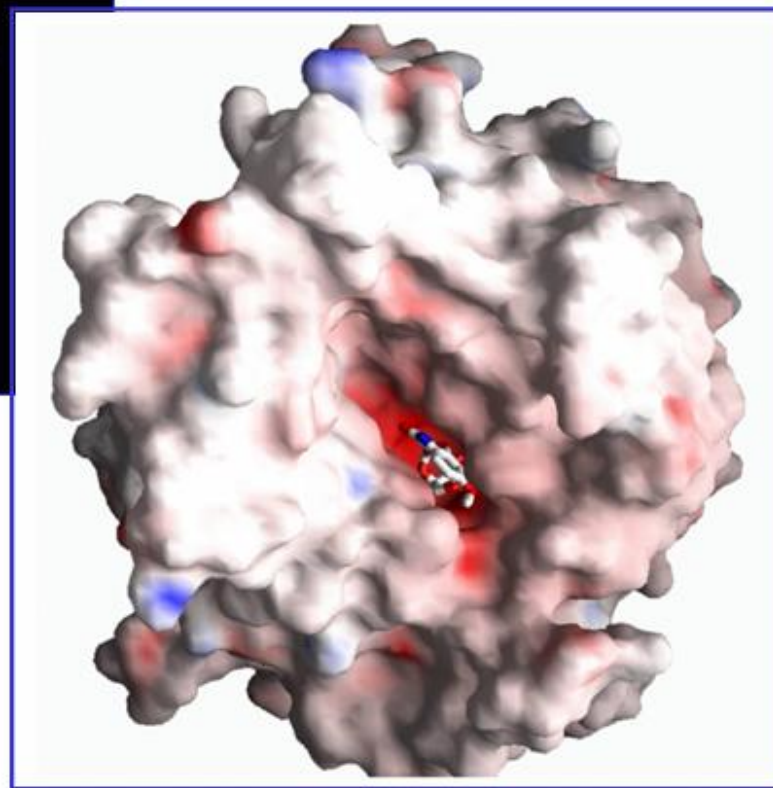
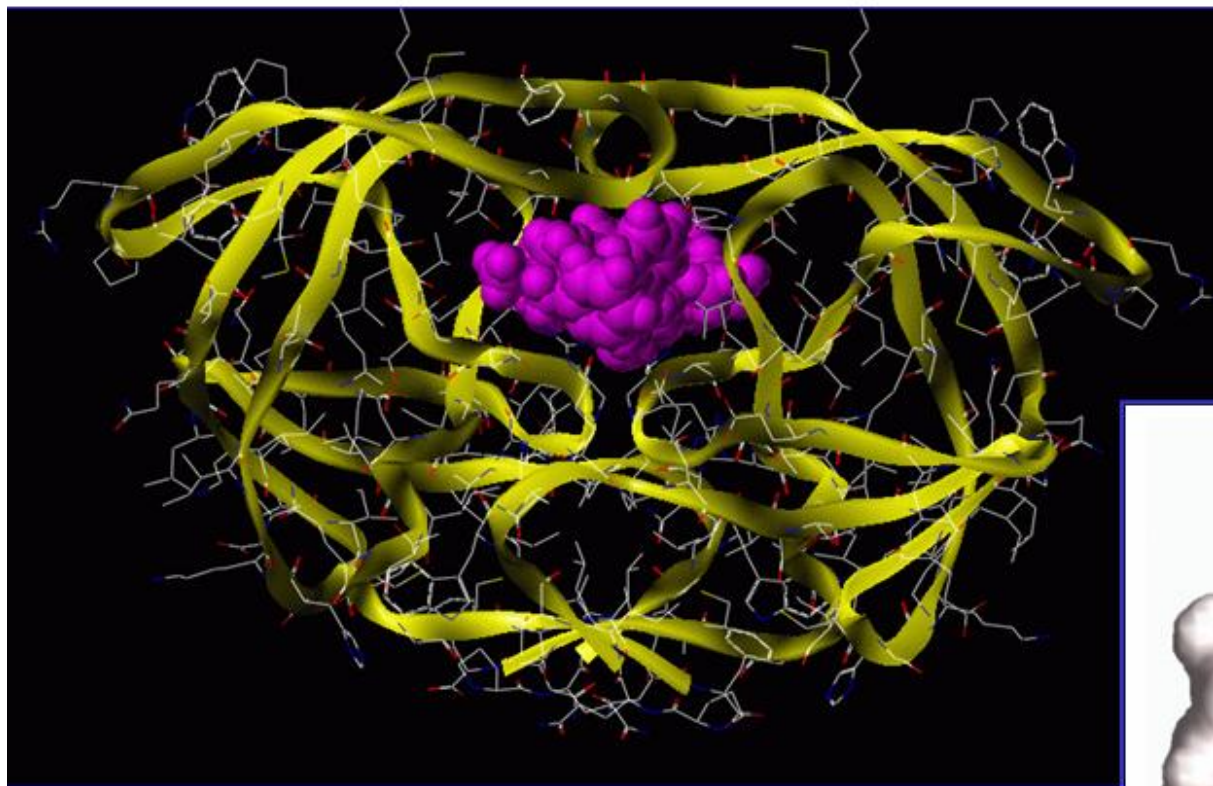


Positional
normalization
removes most
of the
artifact.

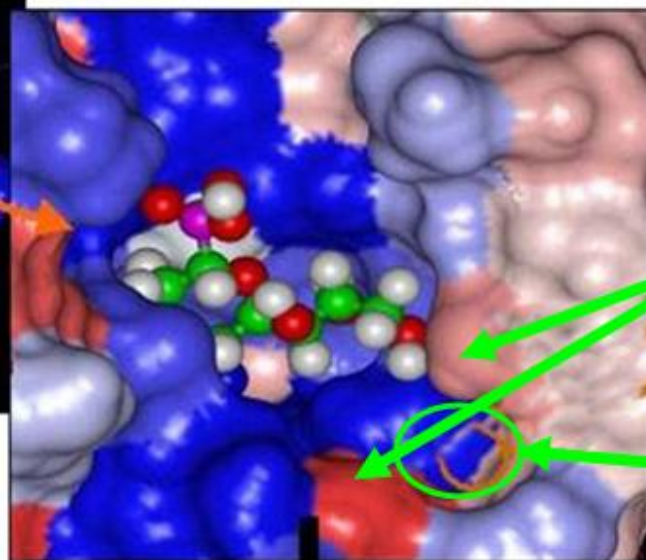
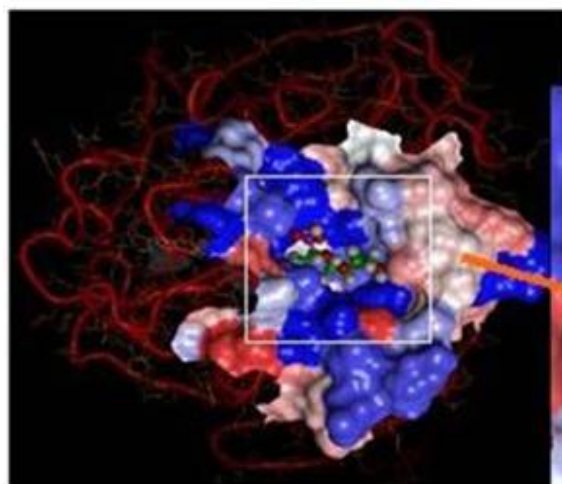
Discovering Drugs Concept 7.5

DRUG BINDING

Drugs bind to proteins by lock and key fit.

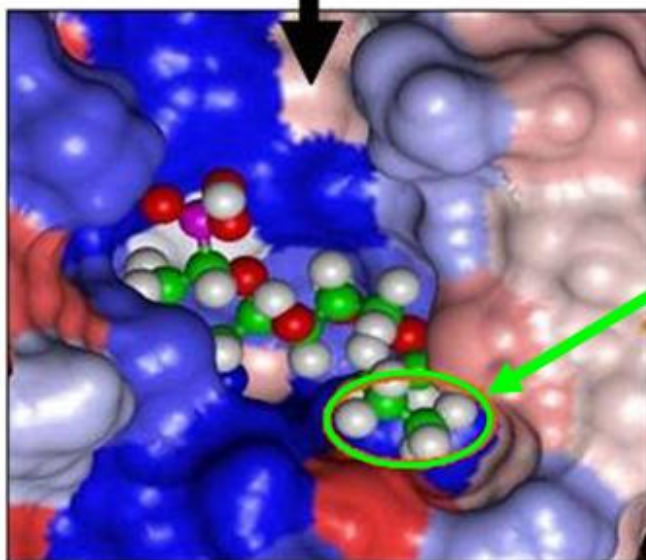


STRUCTURE BASED DRUG DESIGN



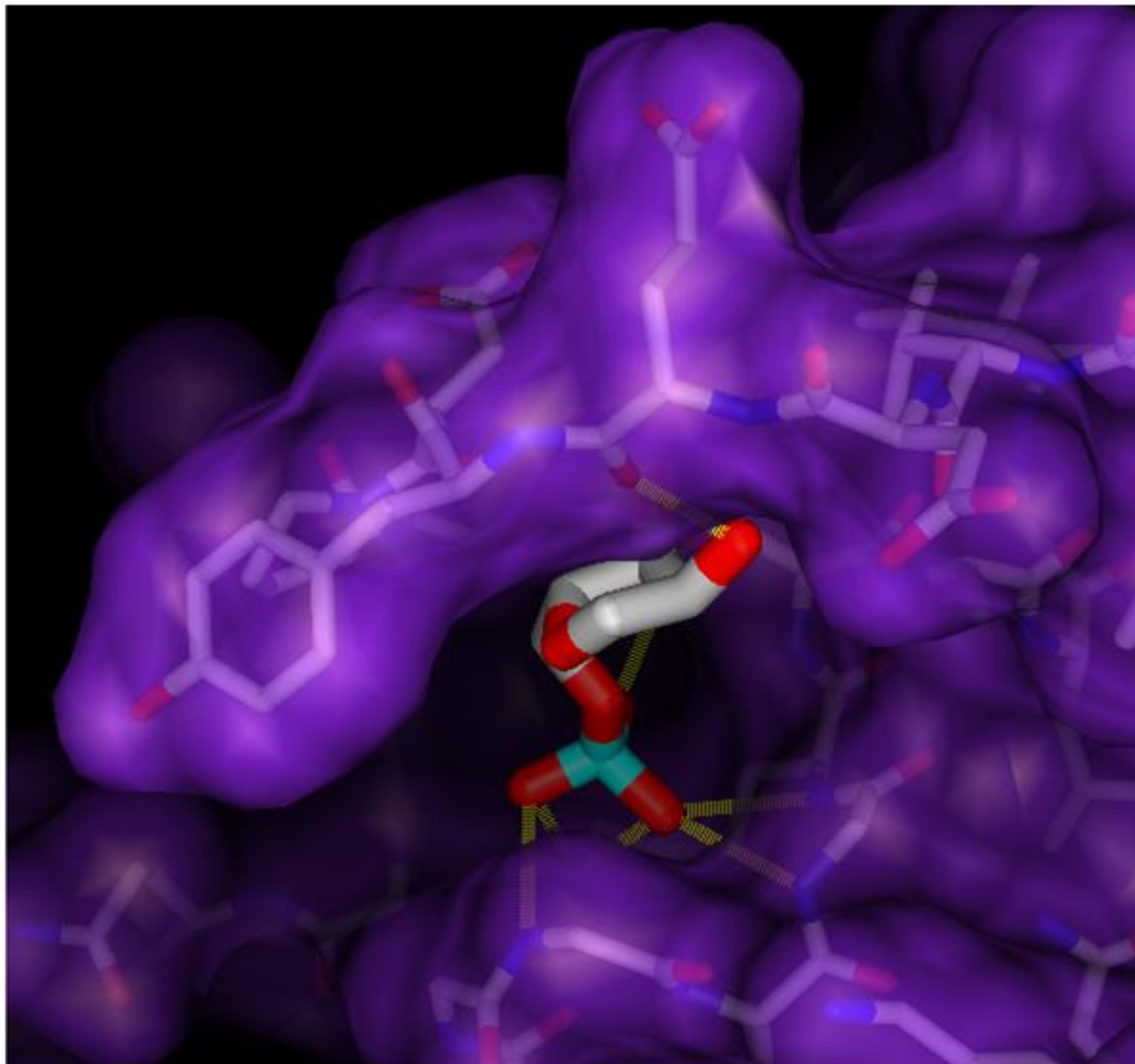
Hydrophobic residue.

Hydrophobic
Pocket



Design drug
to fit pocket

DRUG DOCKING



Kuntz et al. J. Mol. Biol. 161: 269, 1982.

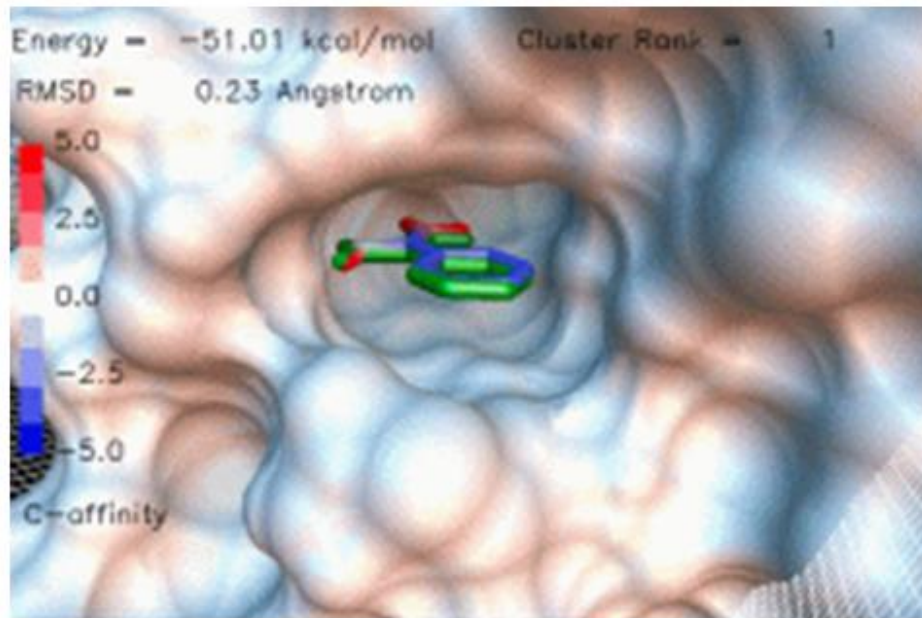
DOCK developed in 1981 by Kuntz and co-workers at UCSF.

Fit into active site by geometrical criteria

<http://dock.compbio.ucsf.edu/>

©Michael Levitt 04

AUTODOCK

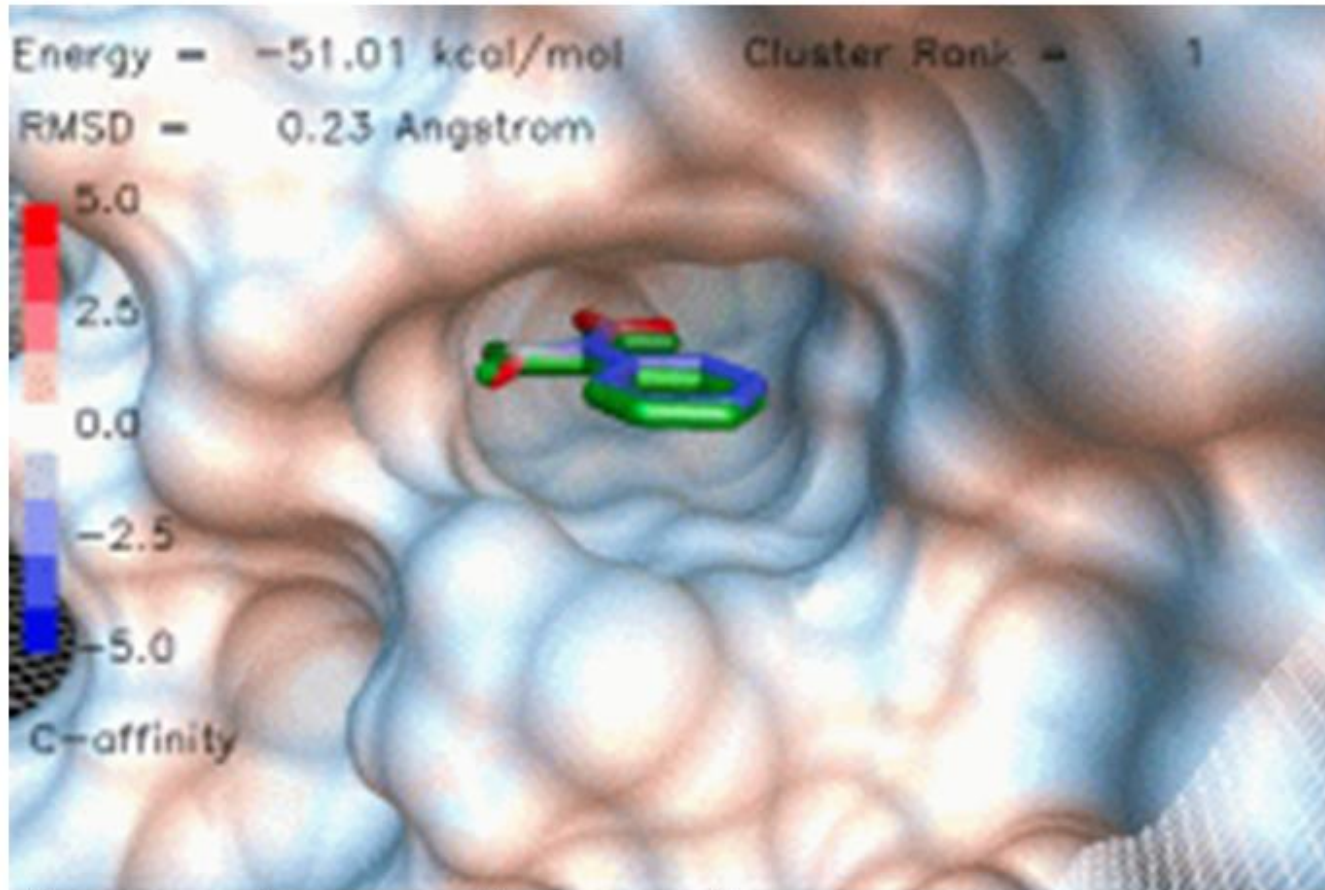


- Use simulated annealing to fit into active site with simple energy function.
- Olson group at Scripps in 1989.

Goodsell & Olson, (1990), Proteins 8: 195-202 (1990)

<http://www.scripps.edu/pub/olson-web/people/gmm/>

AUTODOCK MOVIES



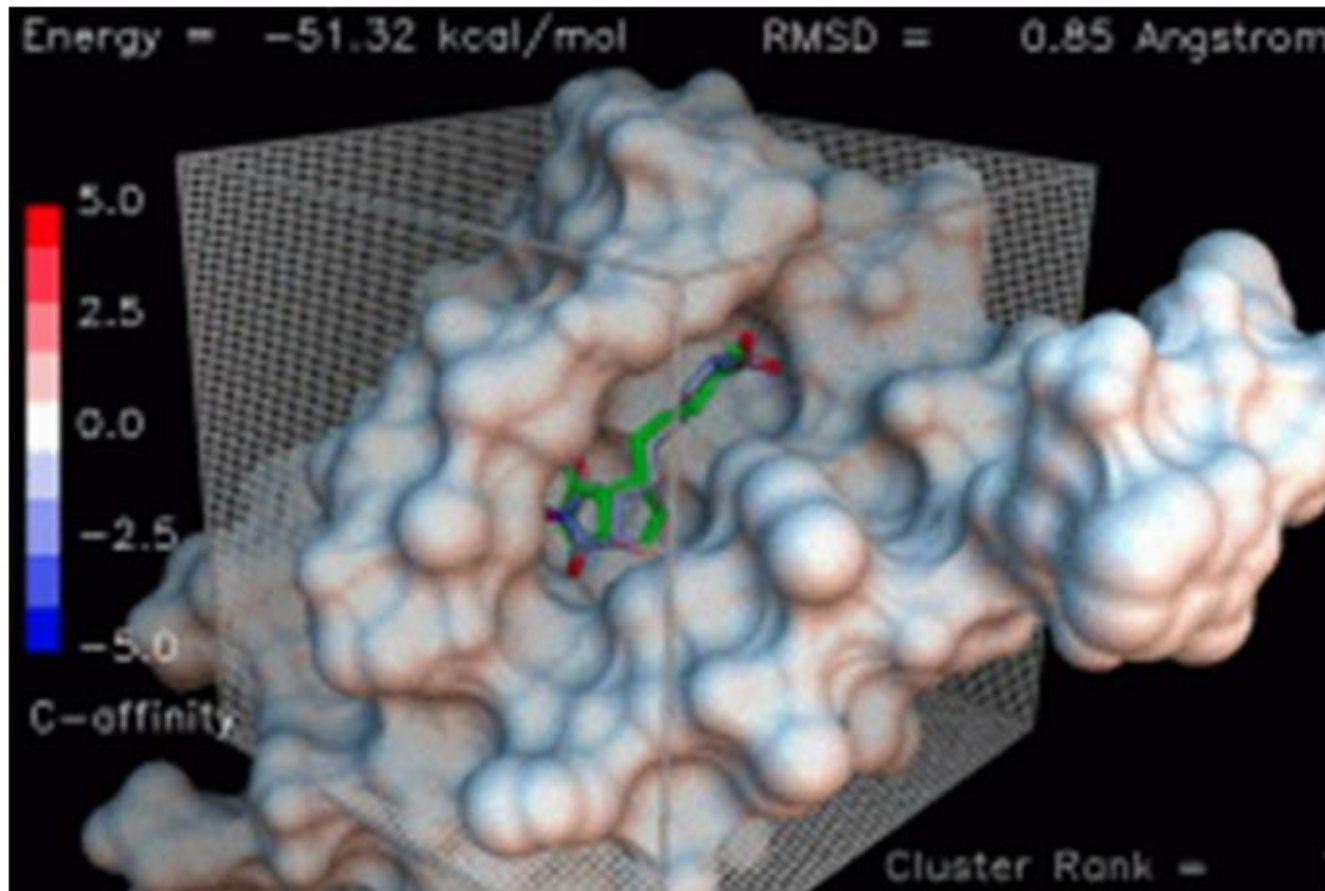
Benzamidine binding to Trypsin.

Gives energies in increasing order.

<http://www.scripps.edu/pub/olson-web/people/gmm/>

©Michael Levitt 04

AUTODOCK MOVIES



Biotin binding to Streptavidin.

Gives energies in increasing order.

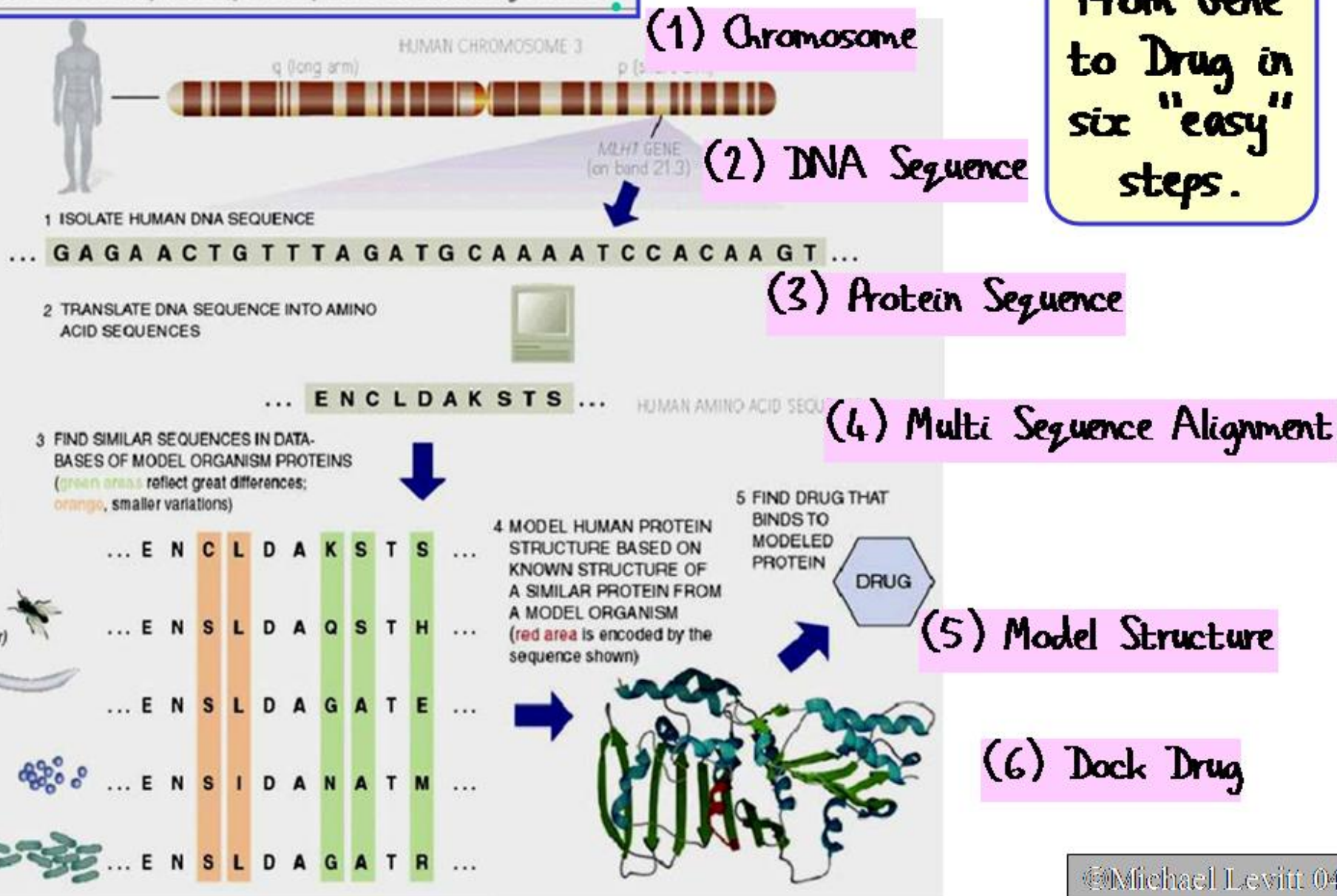
<http://www.scripps.edu/pub/olson-web/people/gmm/>

©Michael Levitt 04

DRUGS FROM SEQUENCE

(c) Mark Gerstein, 1999, Yale, bioinfo.mbb.yale.edu

From Gene to Drug in six "easy" steps.



©Michael Levitt 04

Diagnosing Disease
Concept 7.6

PHARMACOGENOMICS

Little data so far.

PharmGKB
The Pharmacogenetics and Pharmacogenomics Knowledge Base

Search PharmGKB: Go

Home Search Submit Resources Research Network My PharmGKB sign in | help | feedback

Search | **Genes** | Drugs | Diseases

Browse Genes By HGNC Symbol and Name

Display genes with of the following:

- phenotype data
- genotype data
- literature annotations

Go

Jump To:

- All Drugs
- All Diseases
- Browse Index

Legend

- phenotype data available
- genotype data available
- literature annotations available

There are **15,185** genes.

3 1	A 969	B 287	C 1,833	D 529	E 391	F 605	G 670	H 576
I 833	J 25	K 394	L 291	M 973	N 491	O 992	P 1,323	Q 6
R 847	S 1,236	T 1,061	U 208	V 96	W 98	X 42	Y 17	Z 391

Results

- [A1BG](#)
Name: alpha-1-B glycoprotein
- [A2LP](#)
Name: ataxin 2 related protein (non-HGNC gene)
- [A2M](#)
Name: alpha-2-macroglobulin
- [A2MP](#)
Name: alpha-2-macroglobulin pseudogene
- [A4GALT](#)
Name: alpha 1,4-galactosyltransferase
- [AA](#)
Name: atrophy areata, peripapillary chorioretinal degeneration
- [AAAS](#)
Name: achalasia, adrenocortical insufficiency, alacrimia (Allgrove, triple-A)
- [AACP](#)
Name: arylamide acetylase pseudogene
- [AADAC](#)

Lots of genes.

PharmGKB
The Pharmacogenetics and Pharmacogenomics Knowledge Base

Search PharmGKB: Go

Home Search Submit Resources Research Network My PharmGKB sign in | help | feedback

Search | **Genes** | Drugs | Diseases

Browse Genes By HGNC Symbol and Name

Display genes with of the following:

- phenotype data
- genotype data
- literature annotations

Go

Jump To:

- All Drugs
- All Diseases
- Browse Index

Legend

- phenotype data available
- genotype data available
- literature annotations available

There are **182** genes matching your criteria.

3 0	A 30	B 1	C 18	D 6	E 8	F 2	G 10	H 7
I 1	J 0	K 7	L 2	M 5	N 9	O 0	P 8	Q 0
R 4	S 43	T 8	U 10	V 1	W 0	X 2	Y 0	Z 0

Results

- [ABC1](#)
Name: ATP-binding cassette, sub-family B (MDR/TAP), member 1
- [ABC11](#)
Name: ATP-binding cassette, sub-family B (MDR/TAP), member 11
- [ABC4](#)
Name: ATP-binding cassette, sub-family B (MDR/TAP), member 4
- [ABCC1](#)
Name: ATP-binding cassette, sub-family C (CFTR/MRP), member 1
- [ABCC2](#)
Name: ATP-binding cassette, sub-family C (CFTR/MRP), member 2
- [ABCC3](#)
Name: ATP-binding cassette, sub-family C (CFTR/MRP), member 3
- [ABCC4](#)
Name: ATP-binding cassette, sub-family C (CFTR/MRP), member 4
- [ABCG2](#)
Name: ATP-binding cassette, sub-family G (WHITE), member 2

<http://pharmgkb.org/>

PROTEIN 2-D GELS



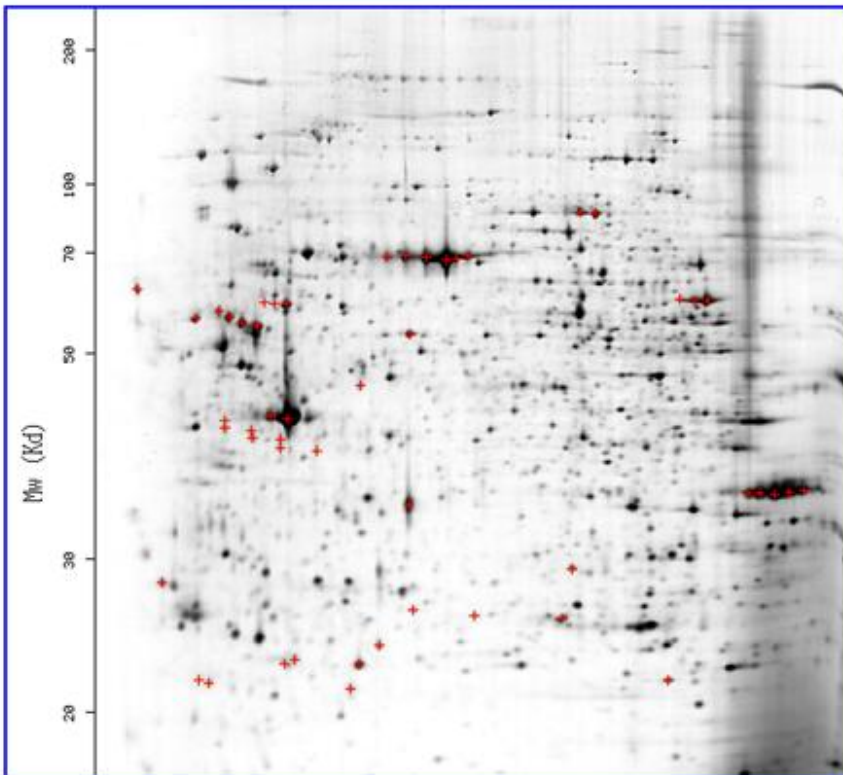
Map Selection: LYMPHOMA_HUMAN

<http://us.expasy.org/ch2d/>

Spots corresponding to known proteins are highlighted in red. Please click on one of them or select the same map in another format:

- [LYMPHOMA_HUMAN large, spots highlighted](#)
- [LYMPHOMA_HUMAN small, spots not highlighted](#)
- [LYMPHOMA_HUMAN large, spots not highlighted](#)

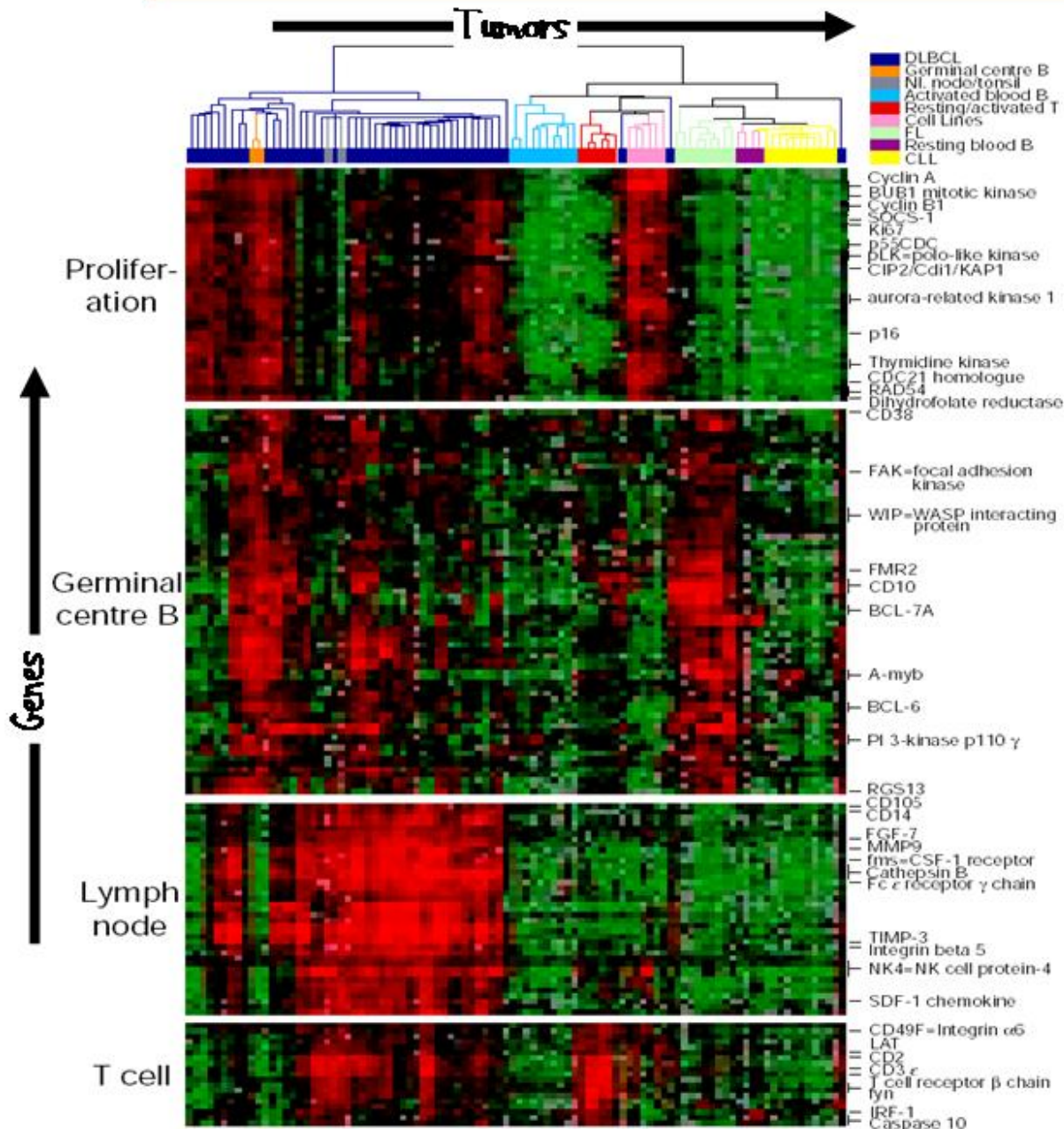
For high resolution gel image, see [reference](#) or [download the tiff file](#).



- Separate proteins in tissue by pI (charge) and Molecular weight (size).
- Need to have a enough protein.
- Look for differences relative to normal tissue.

©Michael Levitt 04

EXPRESSION ARRAY DIAGNOSIS



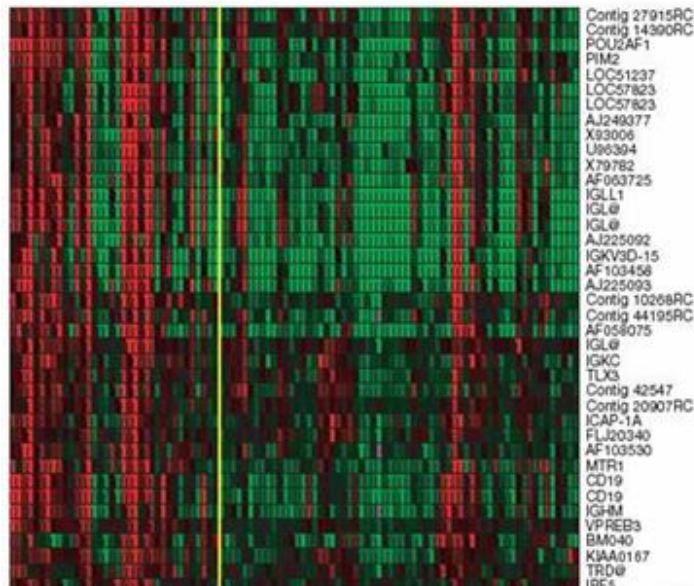
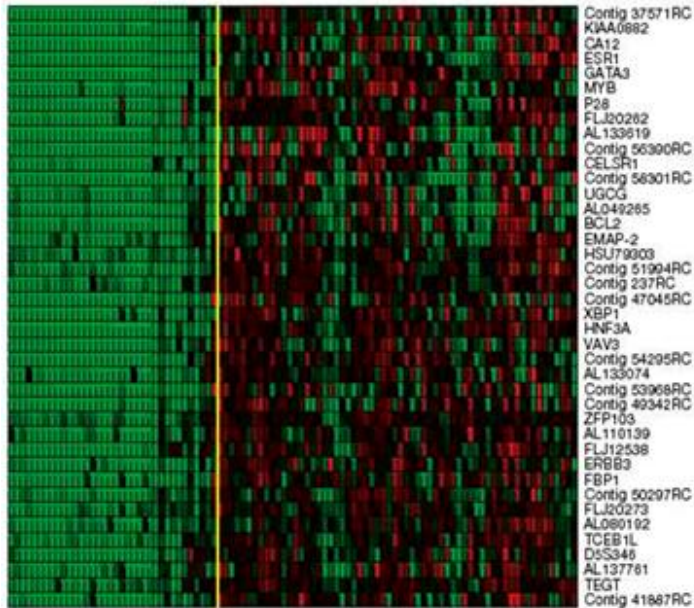
- Cluster genes by similarity of their expression pattern in the different tumors
- Find that this clusters tumors.
- Proliferating cancers are separated from others by the genes that they express.

Red: over-expressed gene in tumor.
Green: under-expressed in tumor.

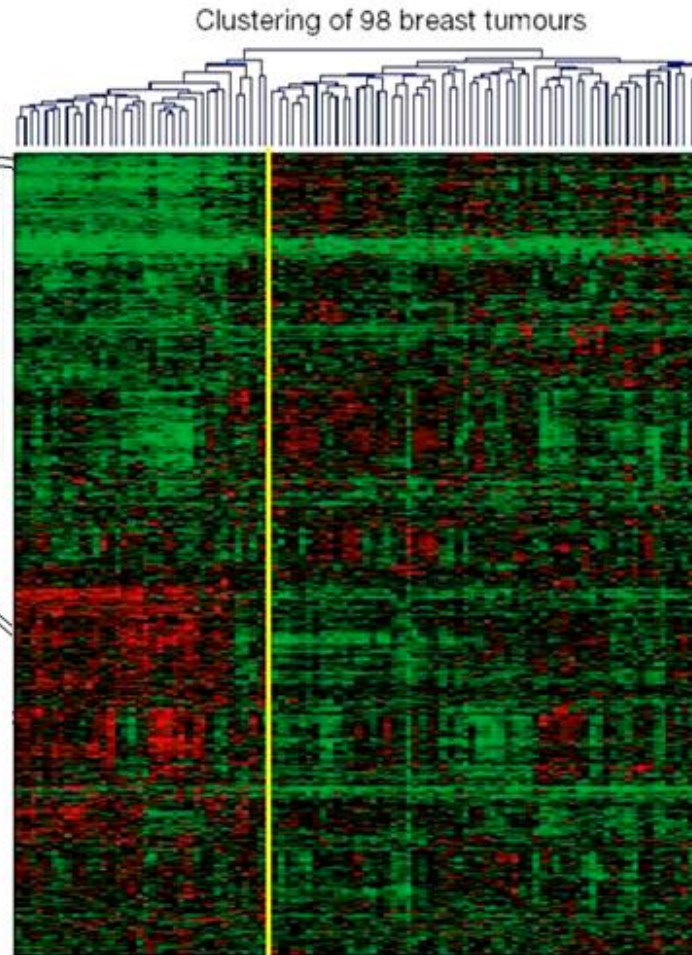
Alizadeh et al. Nature, 403, 503 (2000).

EXPRESSION ARRAY DIAGNOSIS

- Cluster genes by similarity of expression pattern.
- Find that this clusters tumors.



Genes



Clustering of 98 breast tumours

Tumors

Van't Veer et al. Nature, 415, 530 (2002).

Red: over-expressed gene in tumor.
Green: under-expressed in tumor.

Clustering of ~5,000 significant genes