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.

Dozens of papers have been written on this since 1926.

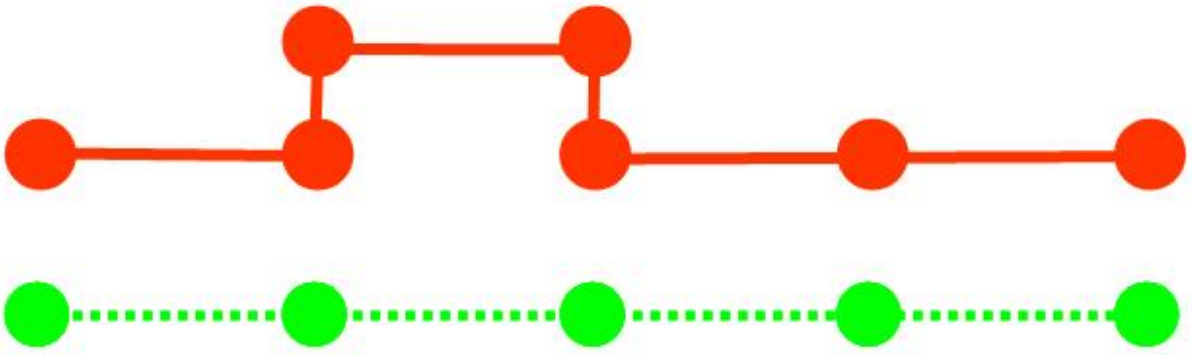
(3) Determine T and t to minimize $\sum (r_A - r'_B)^2$.

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

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

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

SIMPLE STRUCTURAL ALIGNMENT



Molecule A (fixed)

Molecule B

Superimpose B on A #1

Superimpose B on A #2

Superimpose B on A #3

DISTANCE AND SIMILARITY

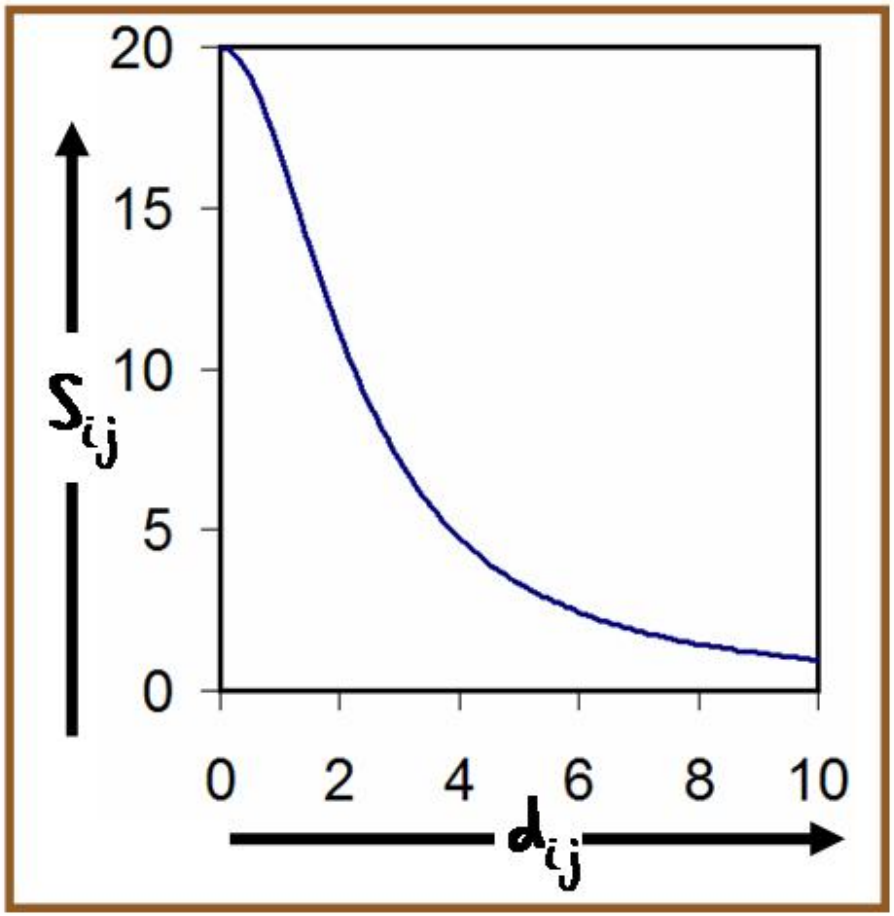
Convert distance between CA atoms I and j to similarity of i and j.

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

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

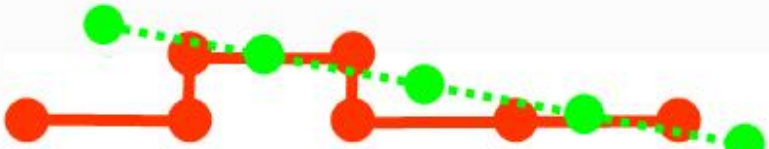
If $d_{ij}^2 = 5$, $S_{ij} = 10$

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

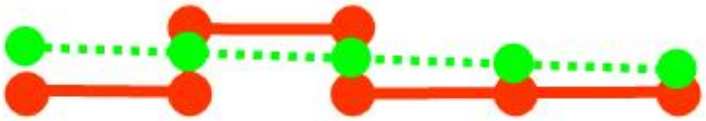


STRUCTAL IN ACTION

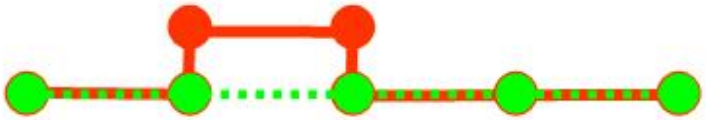
```
- - 1 2 3 4 5      rms  1.96
    | | | | |      brks    0
1 2 3 4 5 6 7      score  56
```



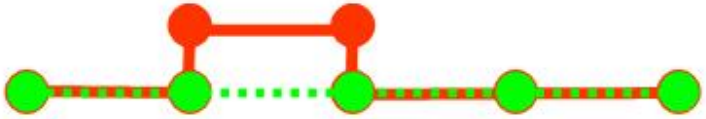
```
1 - 2 - 3 4 5      rms  0.65
|   |   | | |      brks    2
1 2 3 4 5 6 7      score  57
```



```
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
```



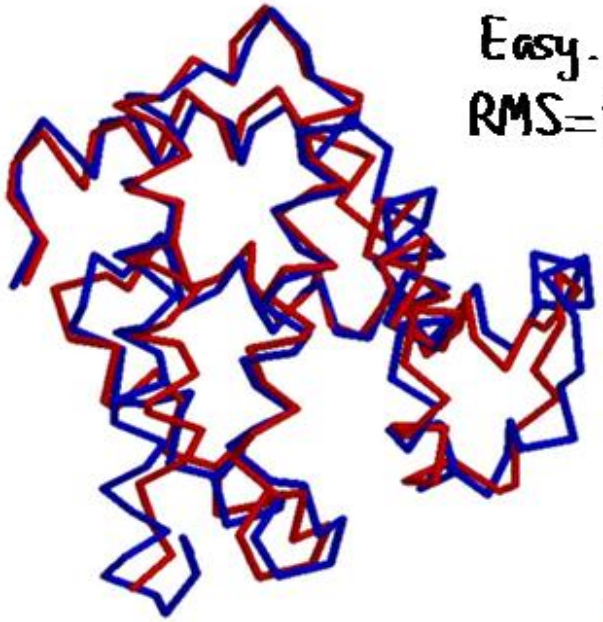
Subbiah et al., Current Biol. 3, 141 (1993).

	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

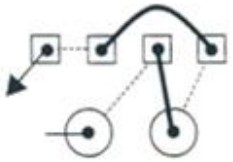
	1	2	3	4	5	6	7
1	19	4	4	1	1	0	0
2	4	16	16	4	4	1	0
3	1	4	4	14	18	4	1
4	0	1	1	4	4	19	4
5	0	0	0	1	1	4	19

	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

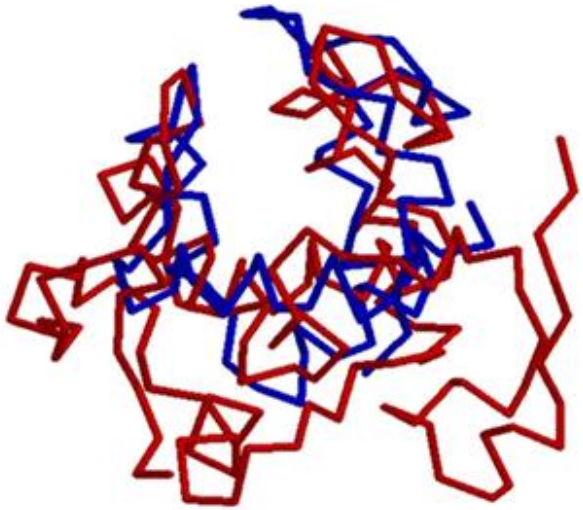
STRUCTURAL ALIGNMENT



Easy. Two globins.
RMS=1.6Å for 135

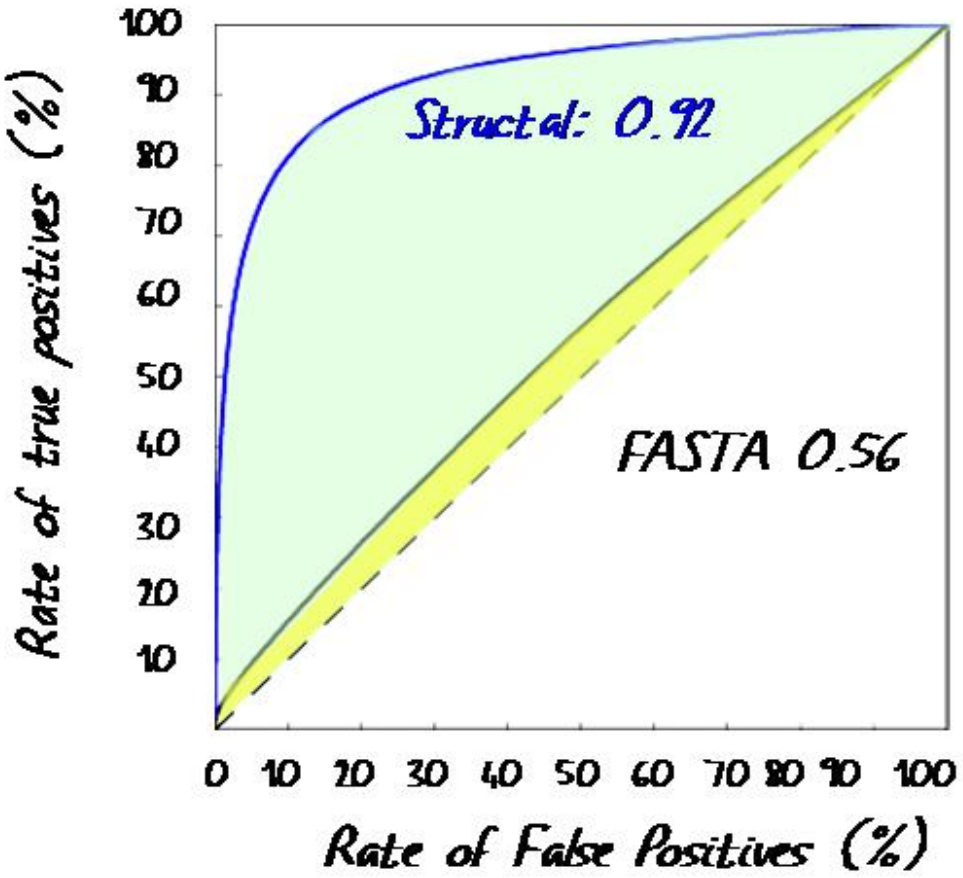


Very Hard.
Same topology
for sub-
domain
(A. Murzin).



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

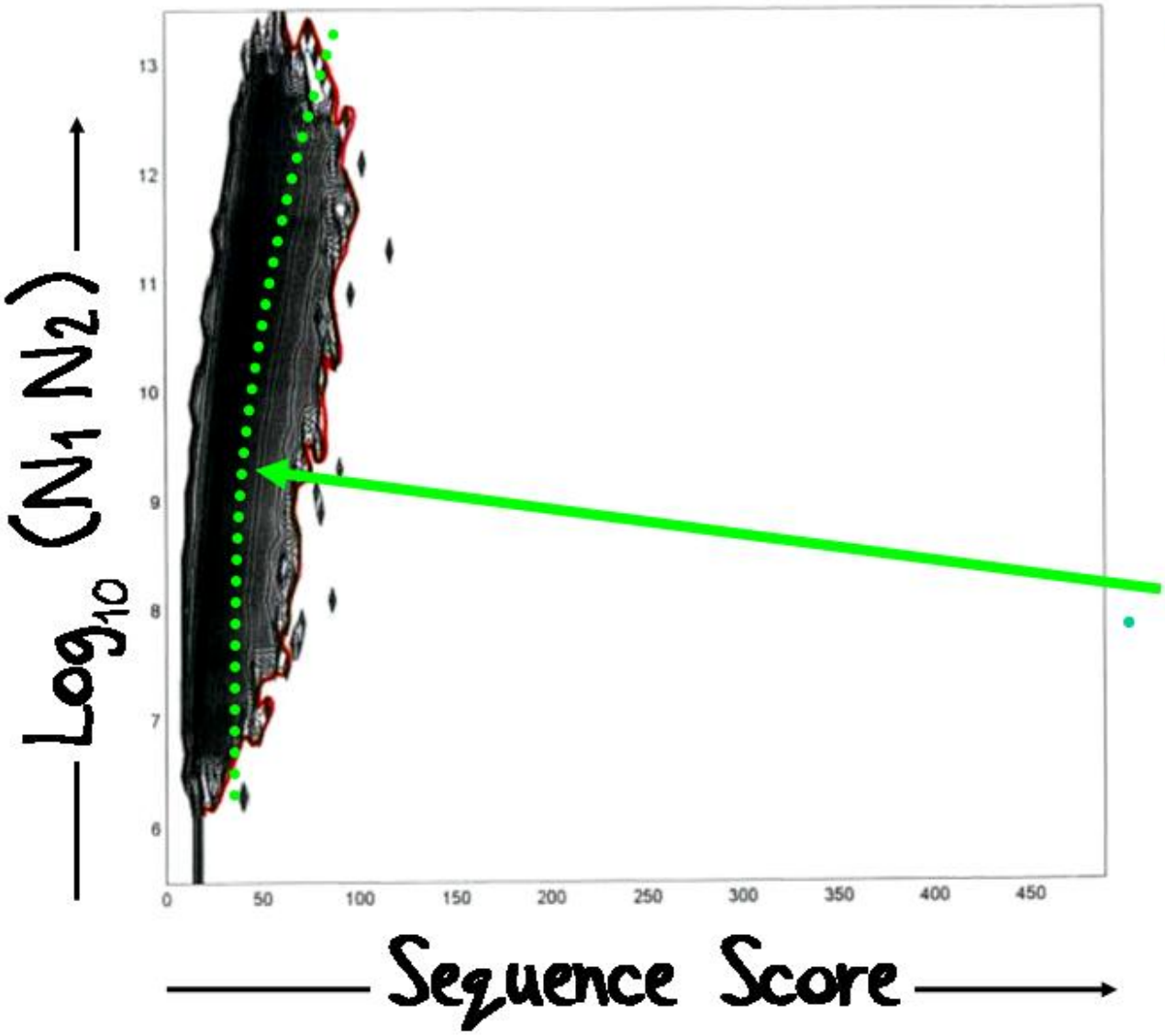
Statistical Significance.

E-values.

Similarity measures.

SEQUENCE SCORE DISTRIBUTION

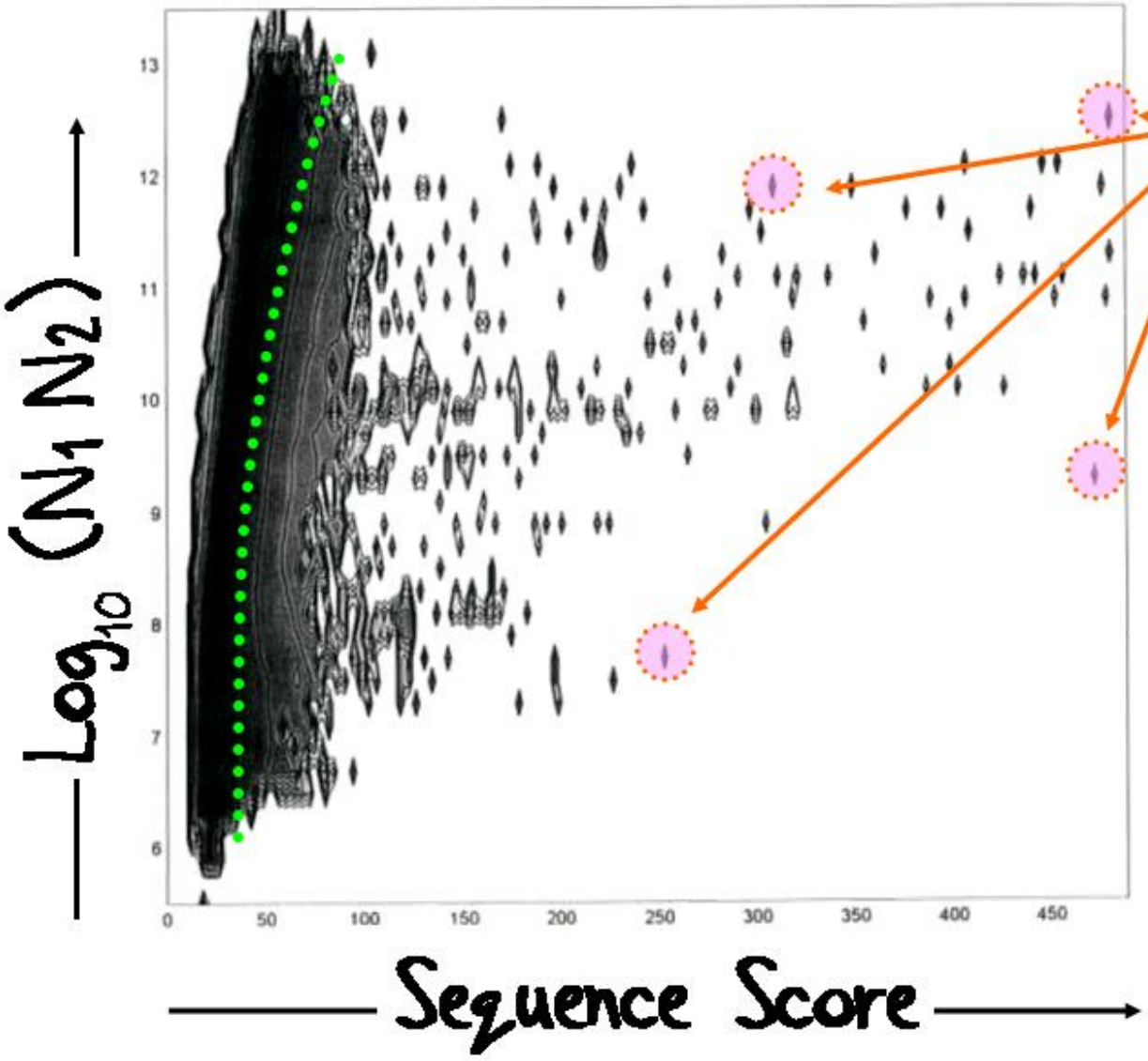
Contour number of occurrences with particular (Score, N) value.



- 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.



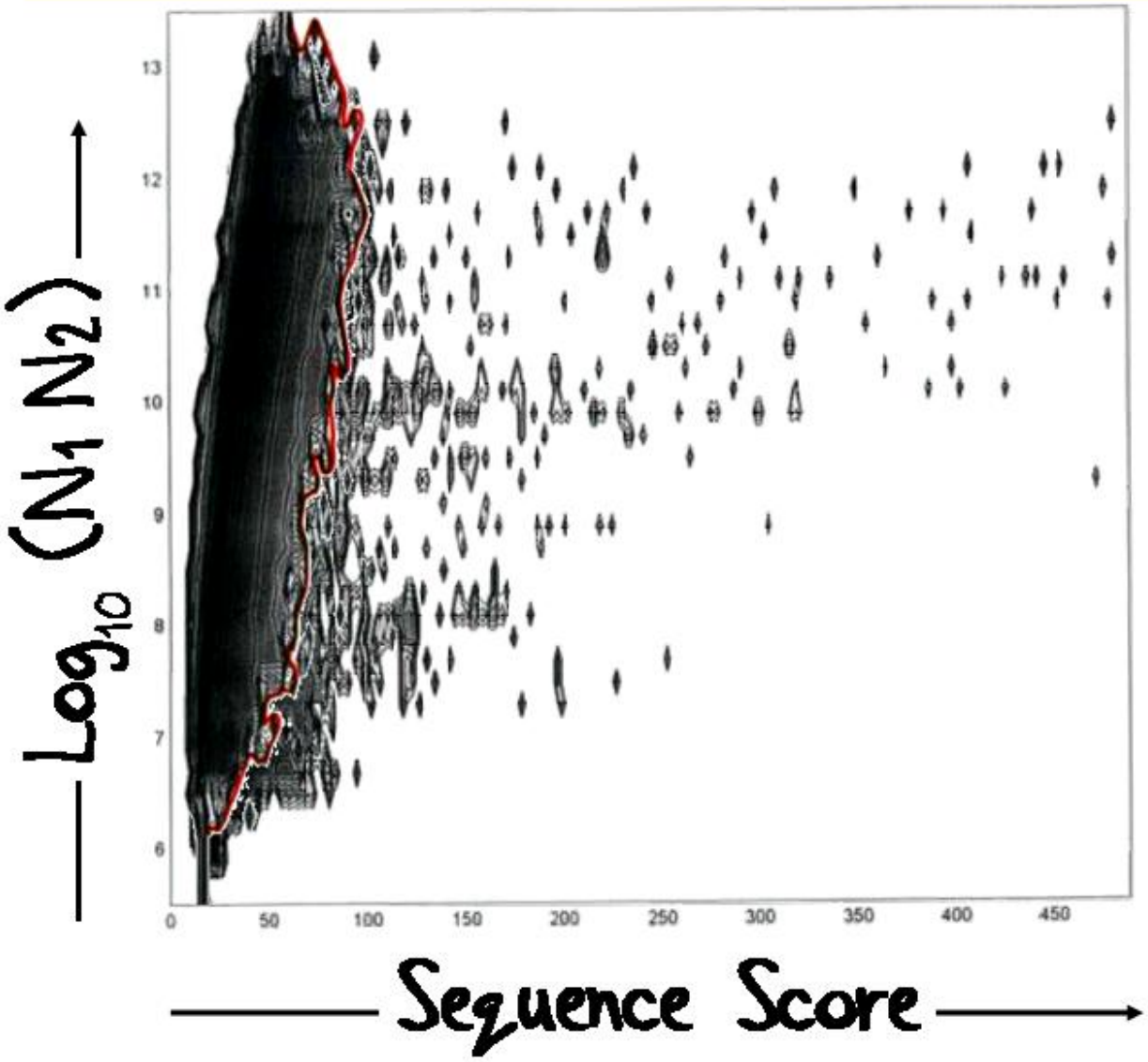
These "islands" are the true positives.

- Include all pairs of sequences.

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

SEQUENCE SCORE DISTRIBUTION

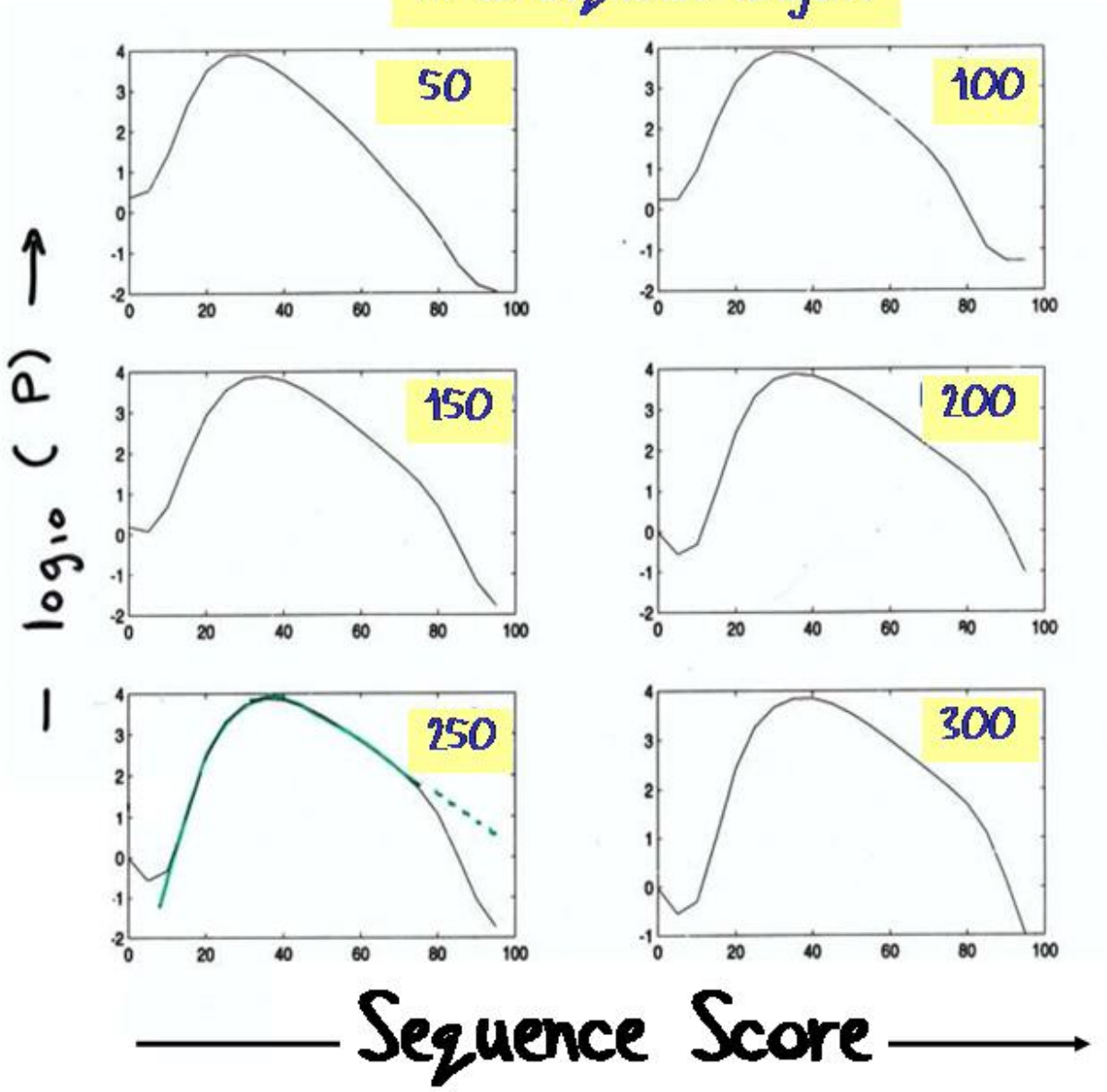
Contour number of occurrences with particular (Score, N) value.



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

SEQUENCE SCORE IS EXTREME VALUE DISTRIBUTION

N is sequence length.

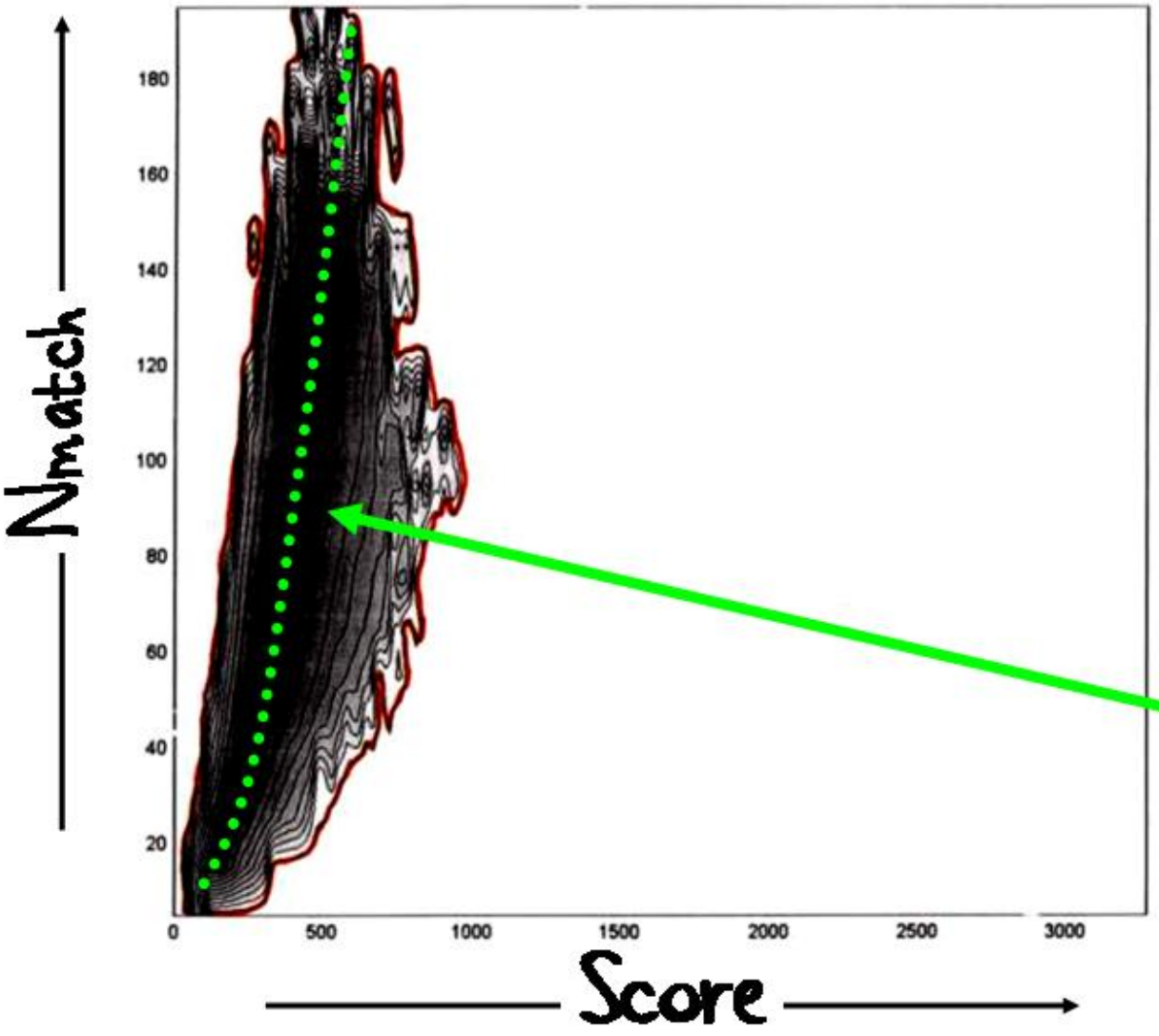


- For the extreme value distribution,
  $P(Z) = \exp(-Z - \exp(-Z))$
  $\log_e P(Z) = -Z - \exp(-Z)$
- For $Z > 0$:
  $\log_e P(Z) = -Z$
- For $Z < 0$:
  $\log_e P(Z) = -\exp(-Z)$.

where $Z = (\text{score} - \text{mean})/SD$.

STRUCTURE SCORE DISTRIBUTION

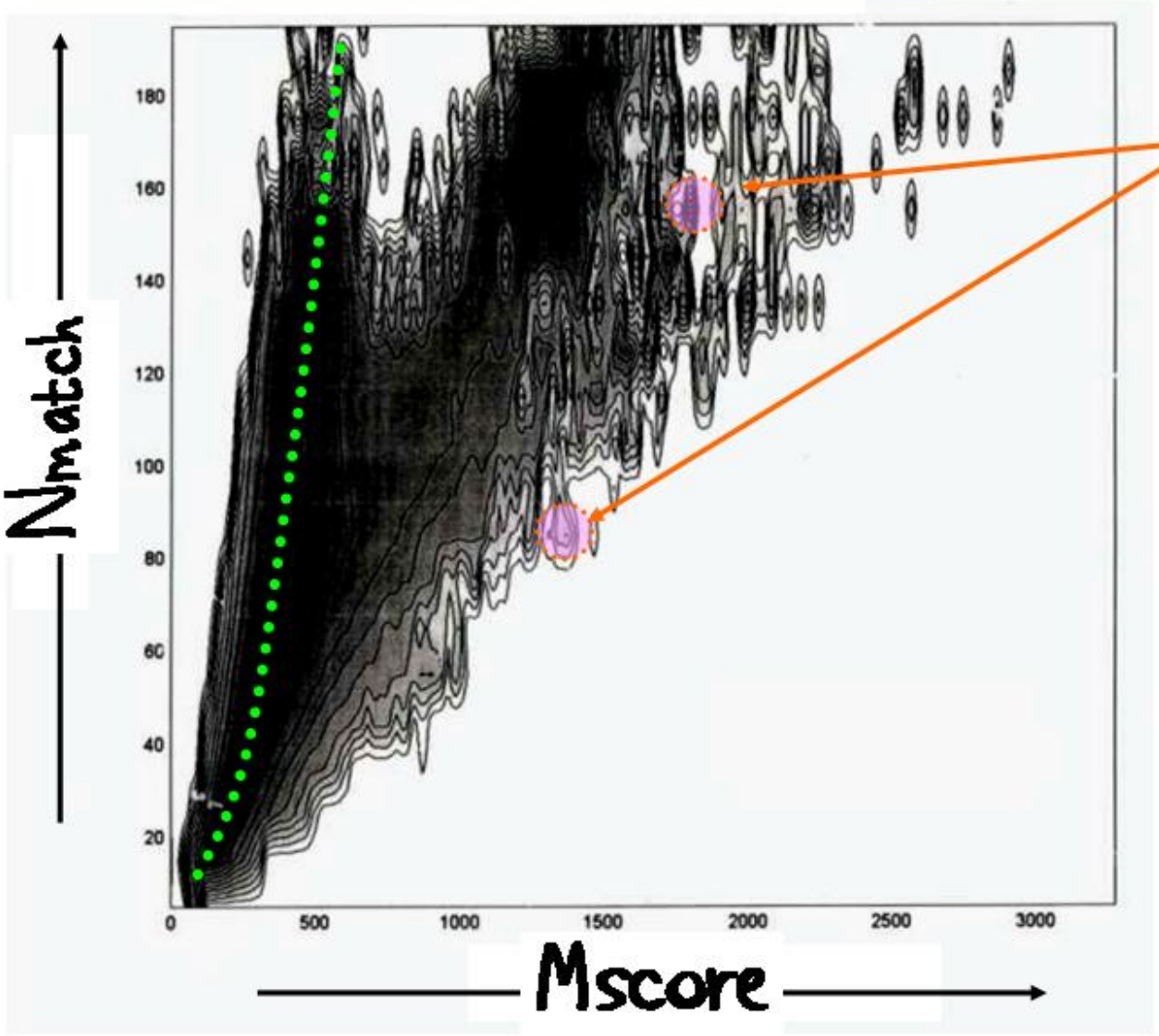
Contour number of occurrences with particular (Score,N) value.



- 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

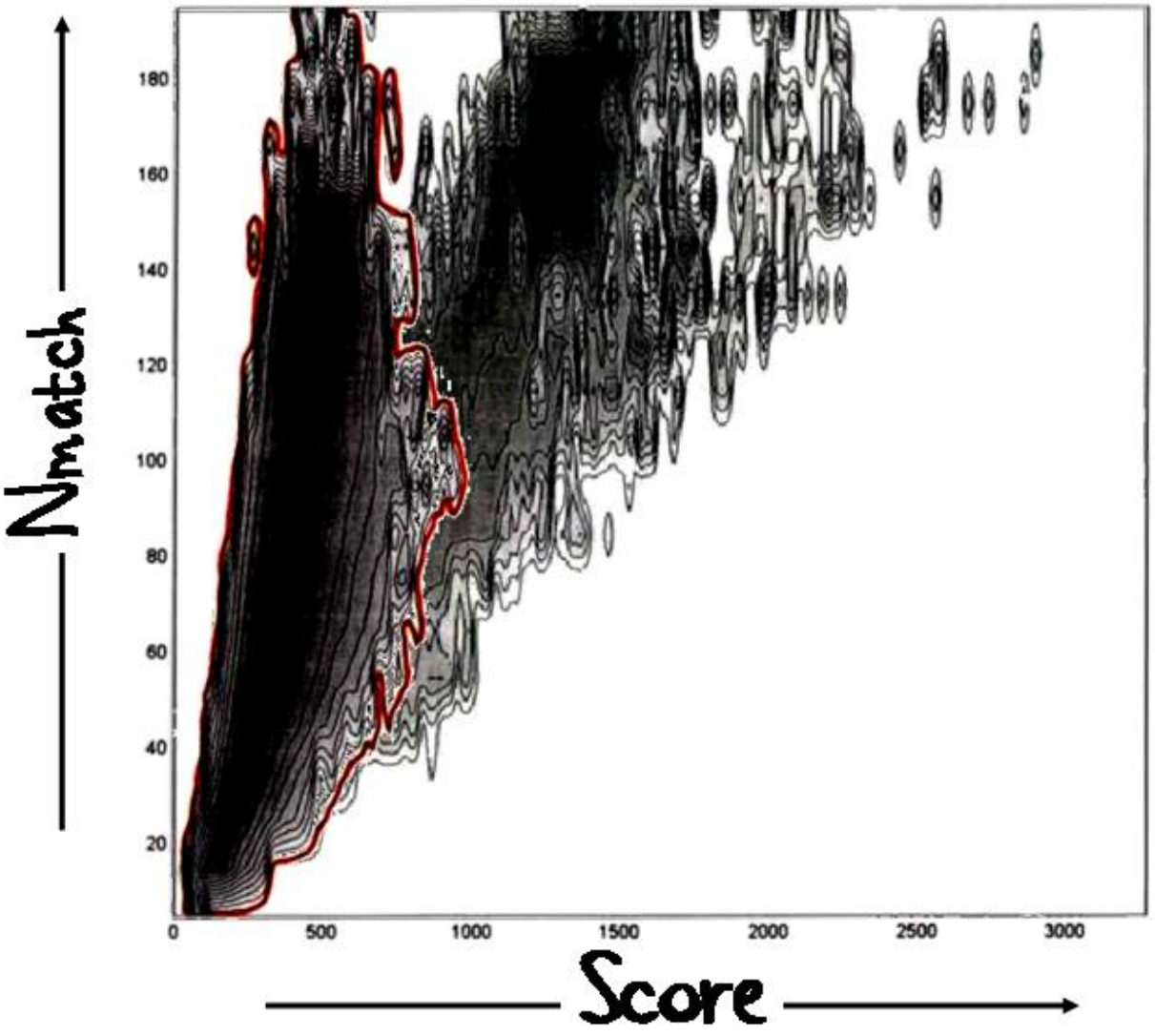
Contour number of occurrences with particular (Score, N) value.



- 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

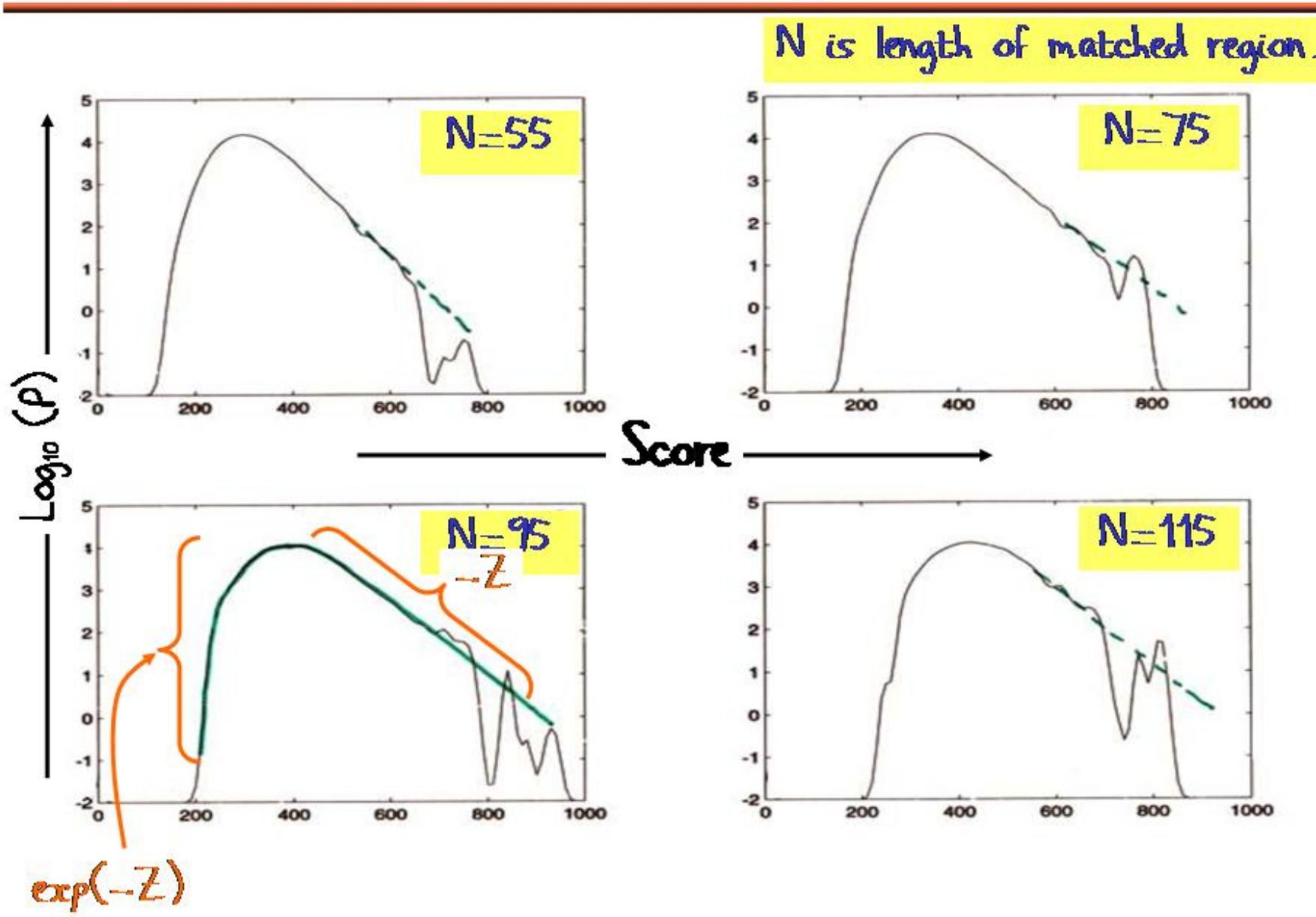
Contour number of occurrences with particular (Score, N) value.



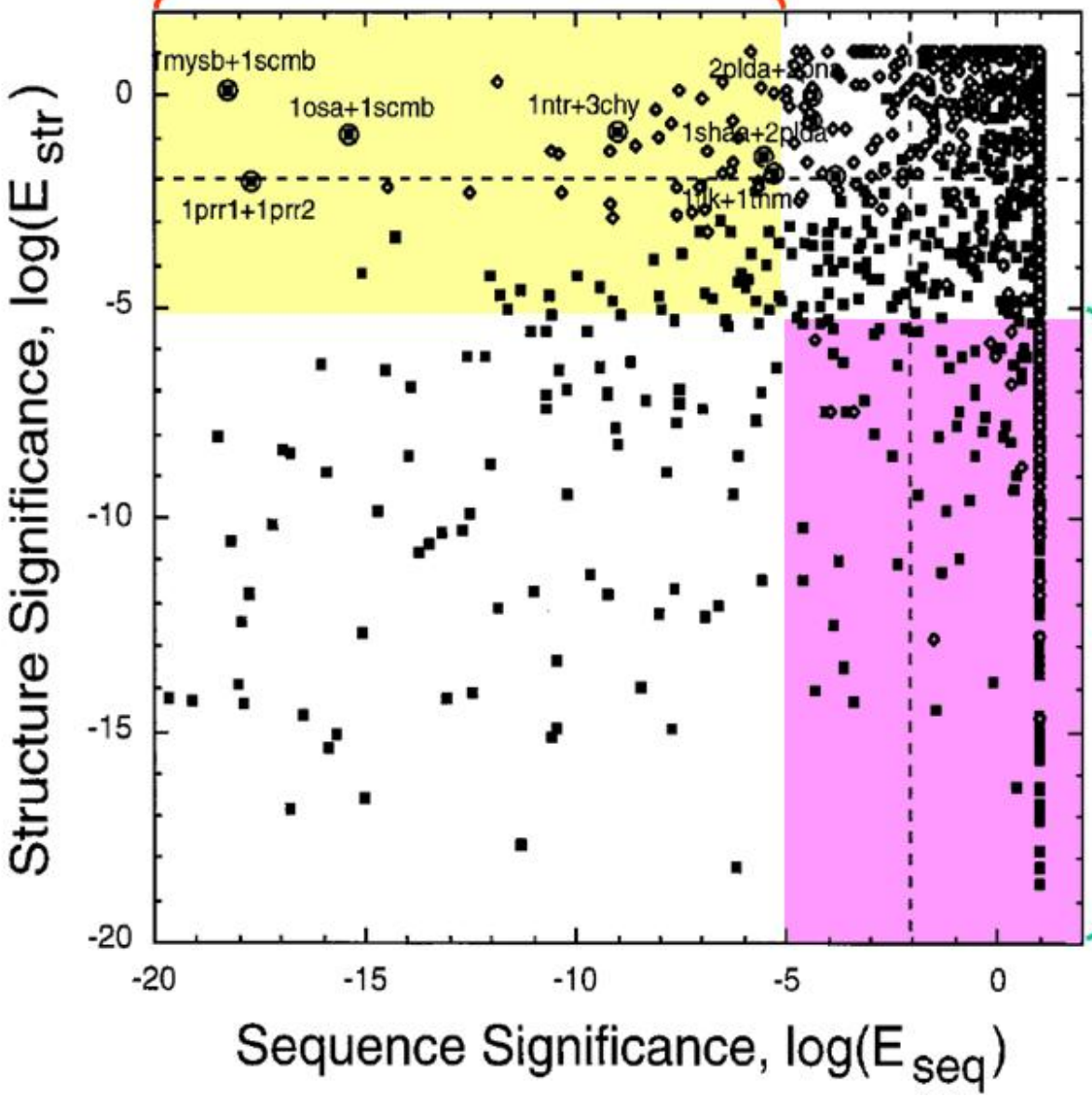
- Show additional pairs in same class.
- Just pairs between SCOP classes.

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.

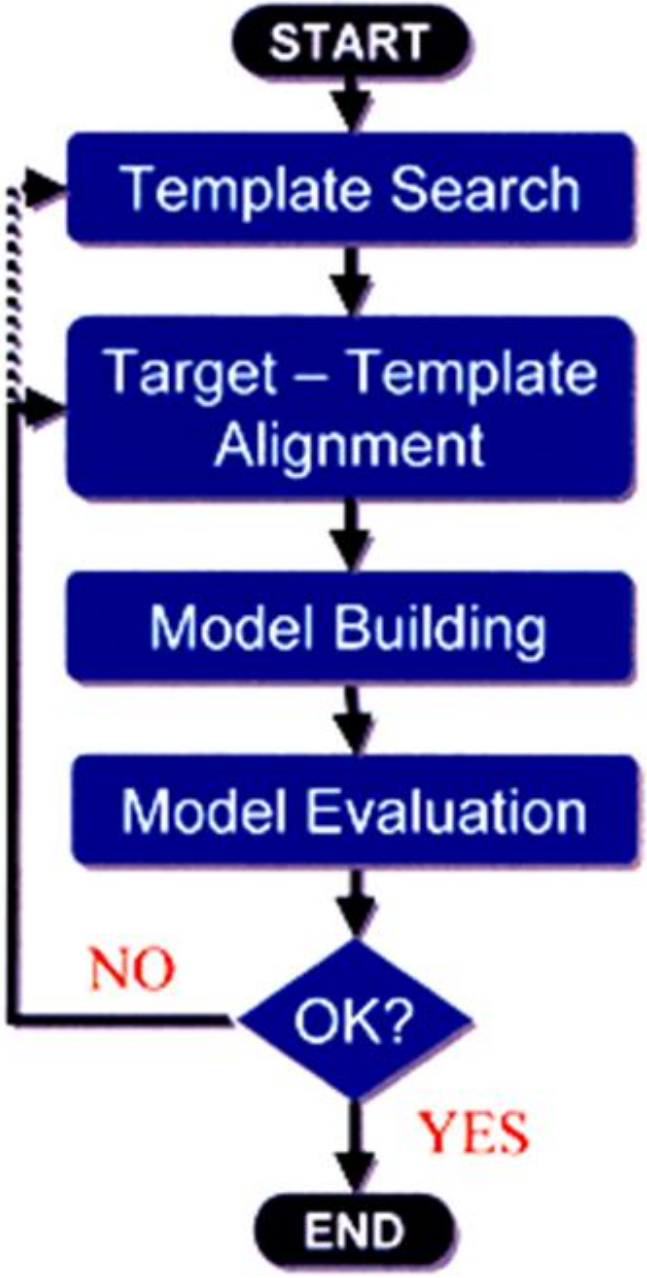
- Neither %ID or RMS are reliable indicators of a significant match.

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



TARGET

TEMPLATE

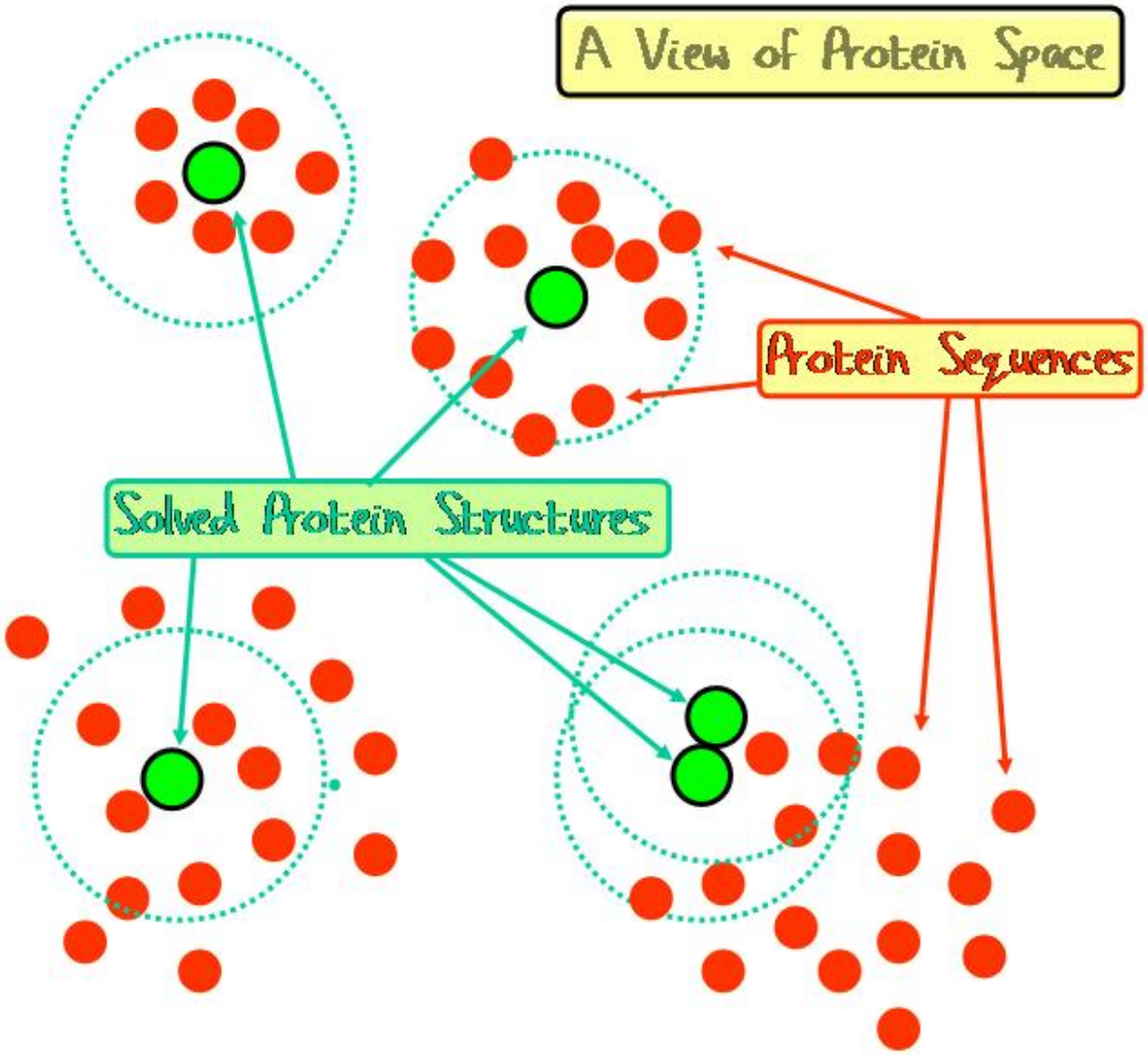
ASILPKRLFGNCEQTSDEGLK
IERTPLVPHISAQNVCLKIDD
VPERLIPERASFQWMNDK

ASILPKRLFGNCEQTSDEGLKIERTPLVPHISAQNVCLKIDDVPERLIPE
MSVIPKRLYGNCEQTSEEAIRIEDSPIV---TADLVCLKIDEIPERLVGE

Comparative or Homology modeling

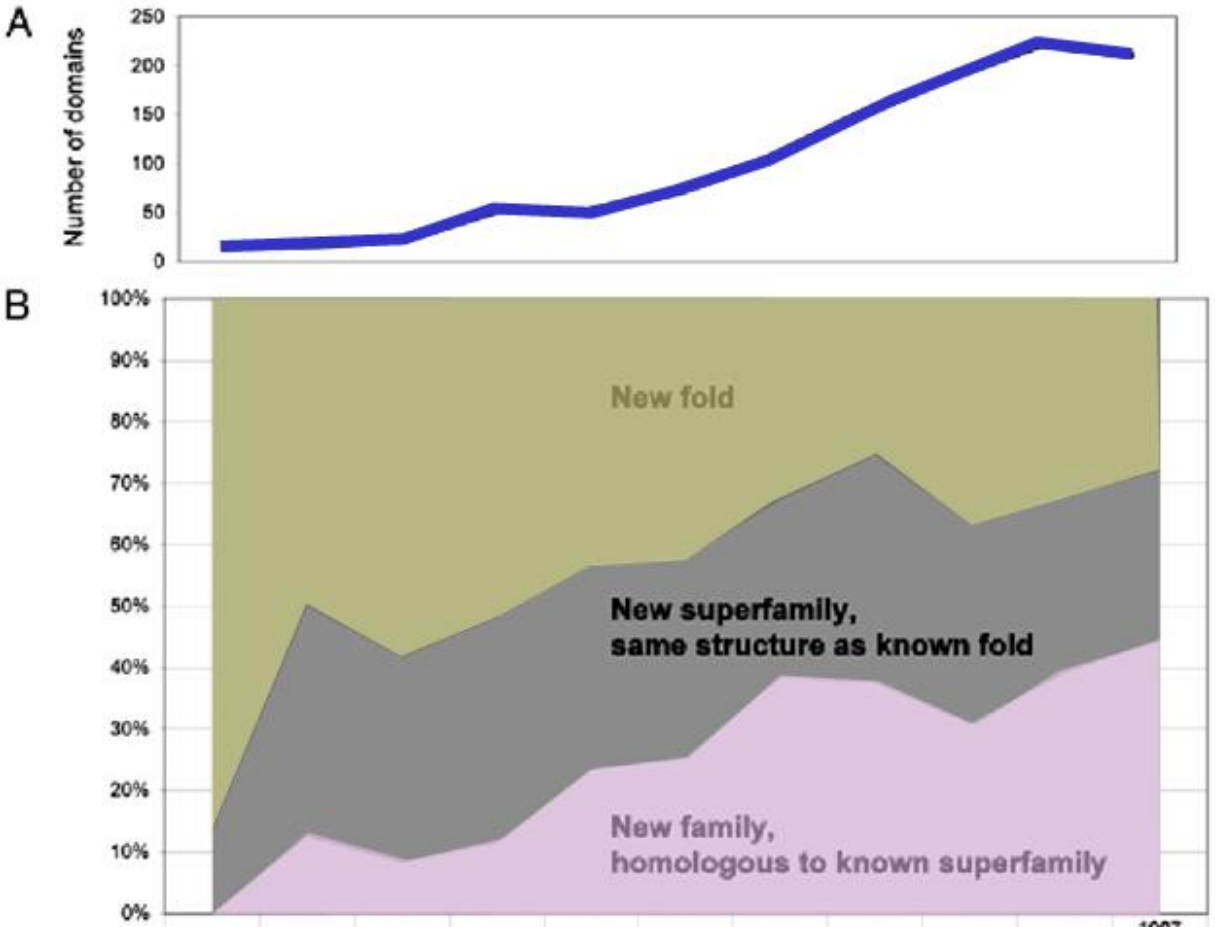
Marti-Rebon, ARBBS, 29, 291 (2000).

STRUCTURAL GENOMICS & MODELING



- 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



	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997 (Partial)
New fold	13	8	14	27	22	30	34	38	69	71	59
New superfamily	2	6	8	19	17	23	30	55	61	64	58
New family	0	2	2	6	12	18	40	56	58	87	94

- 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

- 8 new folds out of 57 deposited PDB files.
- Look at 6303 targets.
- UCSD, Scripps, Stanford.

BSGC CENTER

388 targets.

38 in PDB.

New Folds ?

Progress Summary

BSGC

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

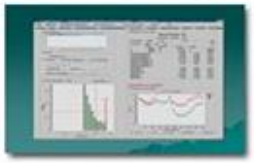
Midwest Center for Structural Genomics

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

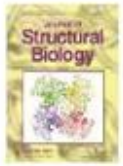
UT Southwestern Medical Center at Dallas

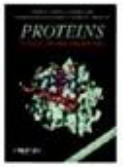
University of Toronto

University of Virginia

1325 targets.
112 in PDB.
15 new folds.

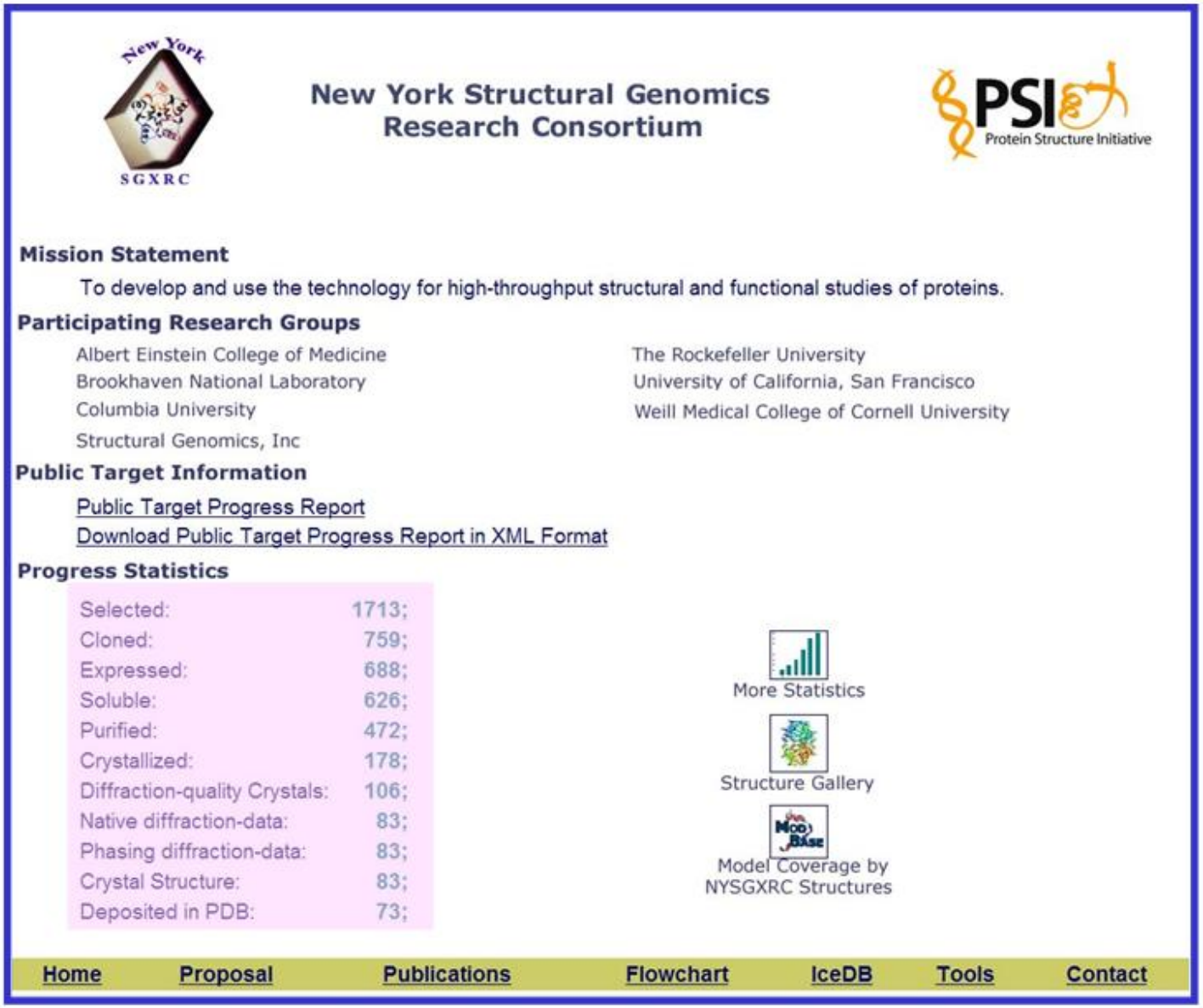
Active Targets: 1325
Crystallized: 318
In PDB: 112
New Folds: 15





The webpage is under construction in W.Minor Lab

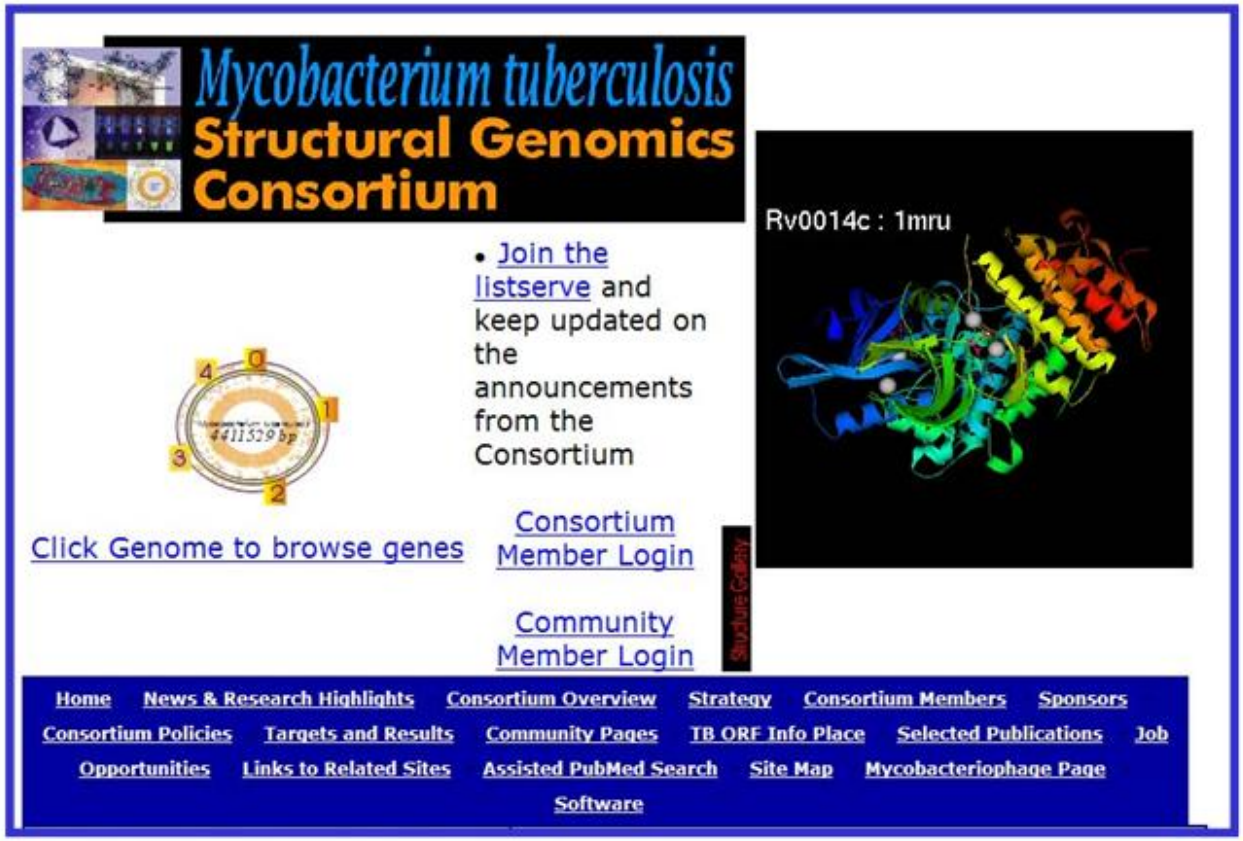
NYSG CENTER



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

1713 targets.
73 in PDB.
? new folds.

TBSG CENTER



- 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

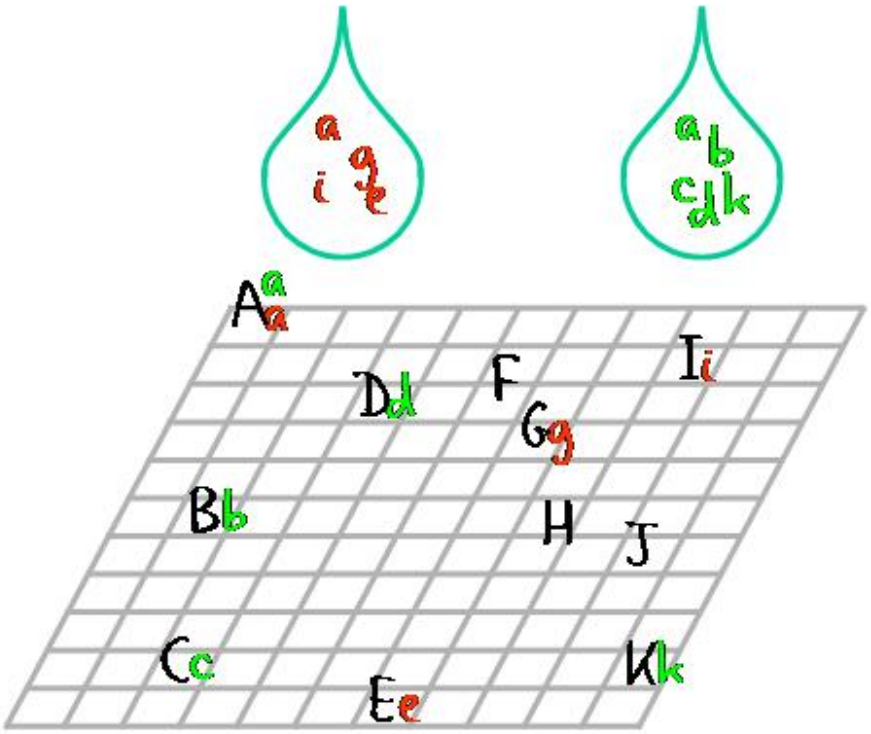
JOINT GENOME	57	8
BERKELEY	38	?
MIDWEST	112	15
NEW YORK	73	?
TUBERCULOSIS	43	?

Estimate 15% new so about 50 new folds.

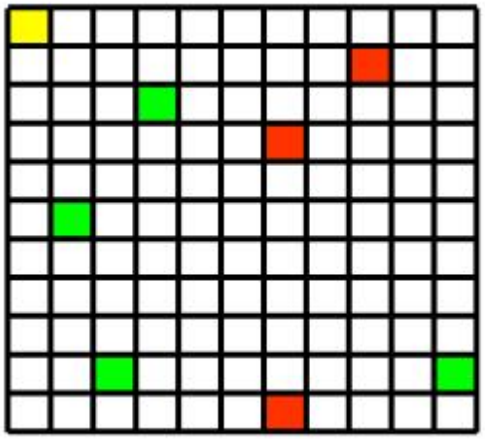
Expression Patterns
Concept 7.4

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 ...).



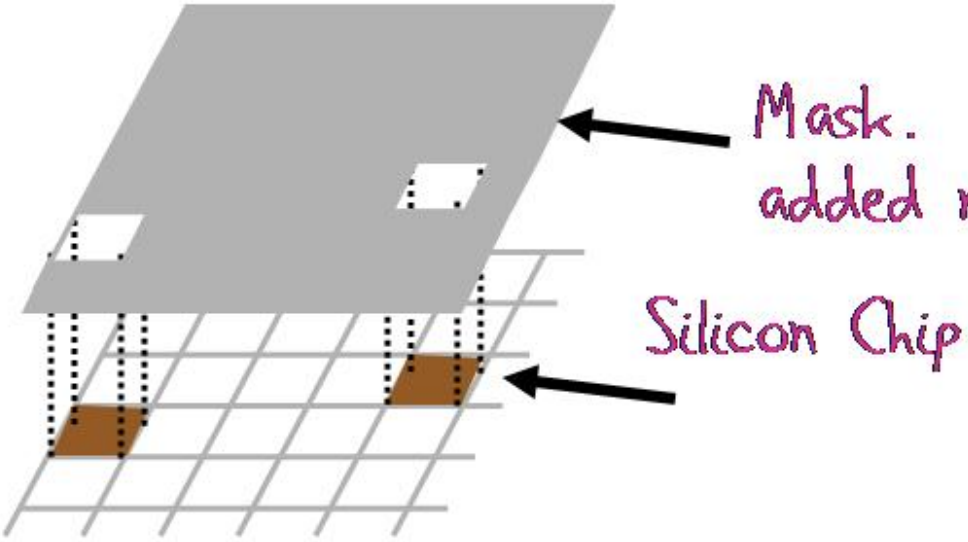
- Assume A only binds a, B only binds b, etc.



- 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 um features). 1,300,000 oligos per chip.
- Need several about 20 oligos per gene.



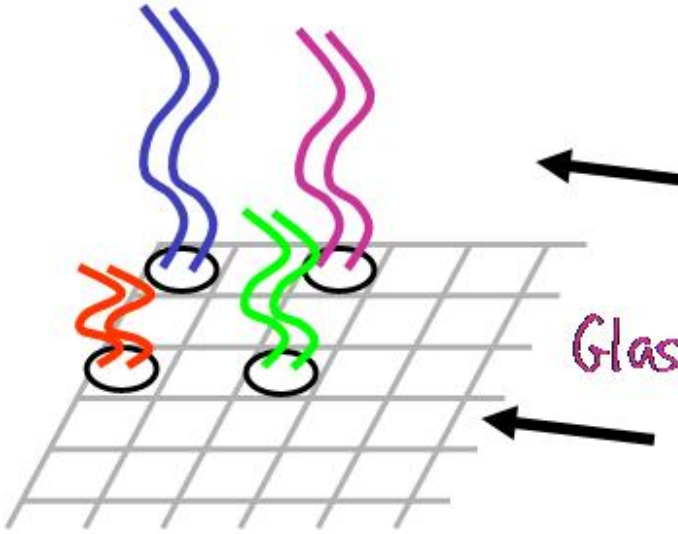
Mask. Need 4 mask for each added nucleotide (A, G, C, T)

Silicon Chip

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.



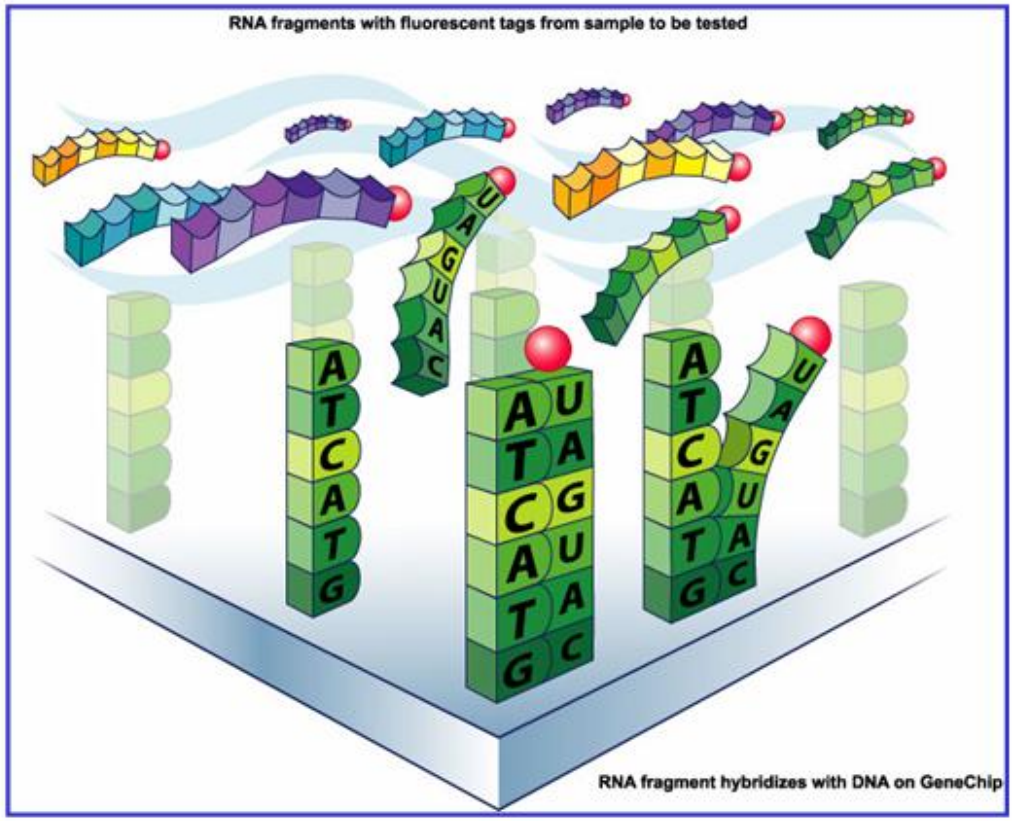
Spot cDNA with robot or ink-jet printer. Make DNA from sequence.

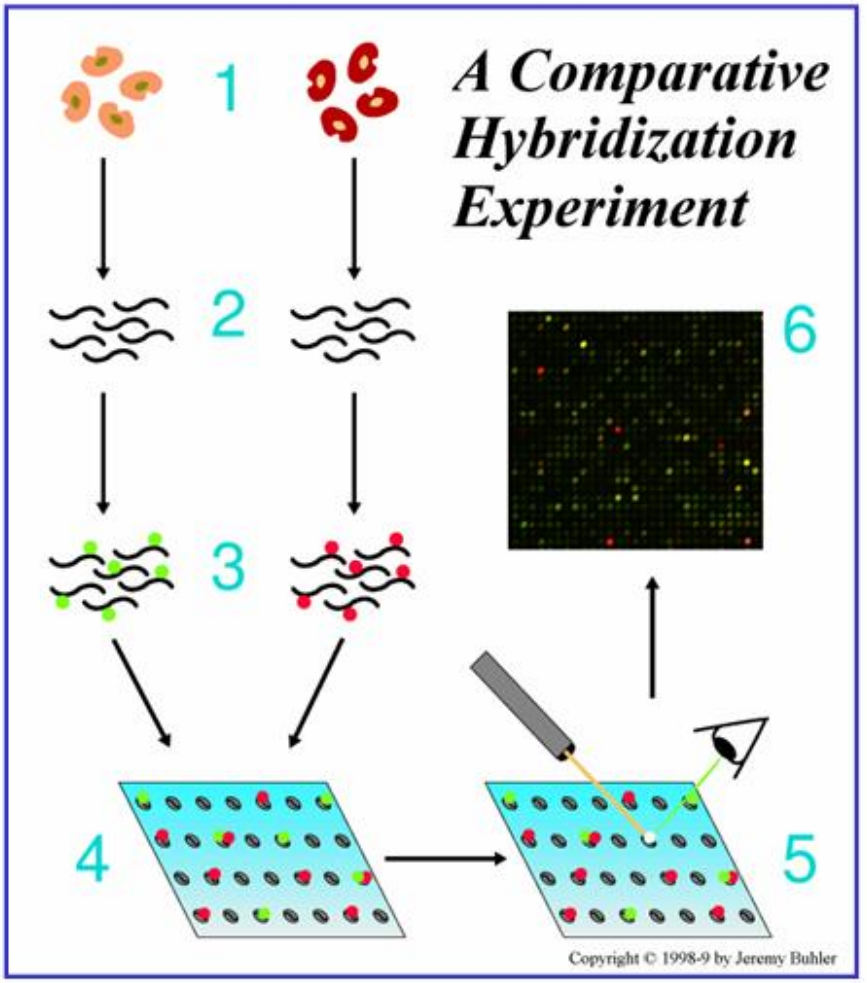
Glass slide

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

MORE EXPRESSION ARRAY CARTOONS

cDNA chip also hybridizes RNA

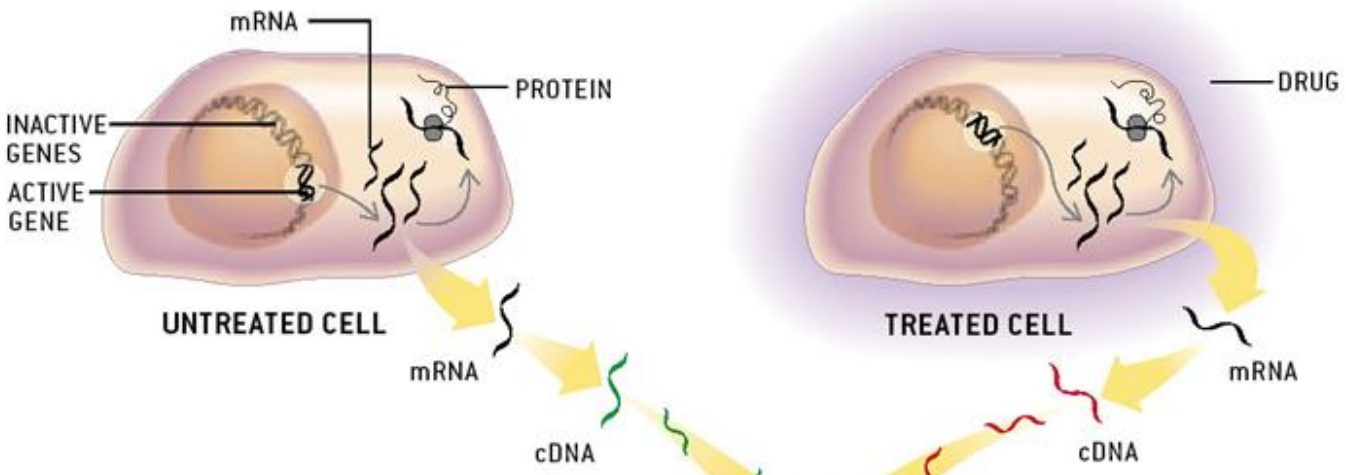




Oligo DNA Chip hybridizes RNA

SCIENTIFIC AMERICAN EXPLAINS

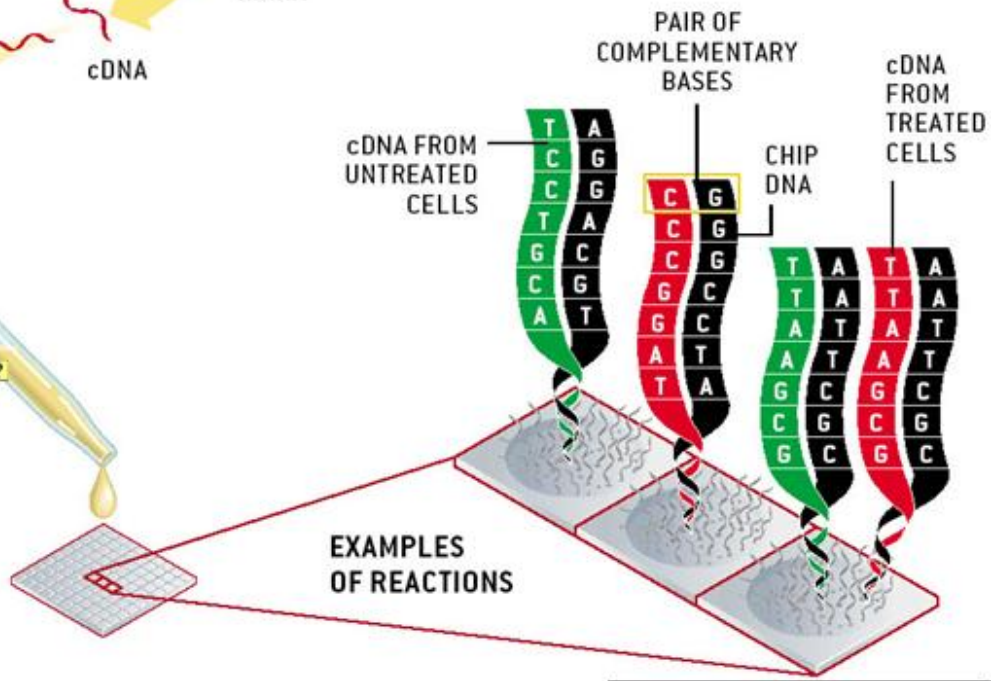
Experimental setup



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.

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.

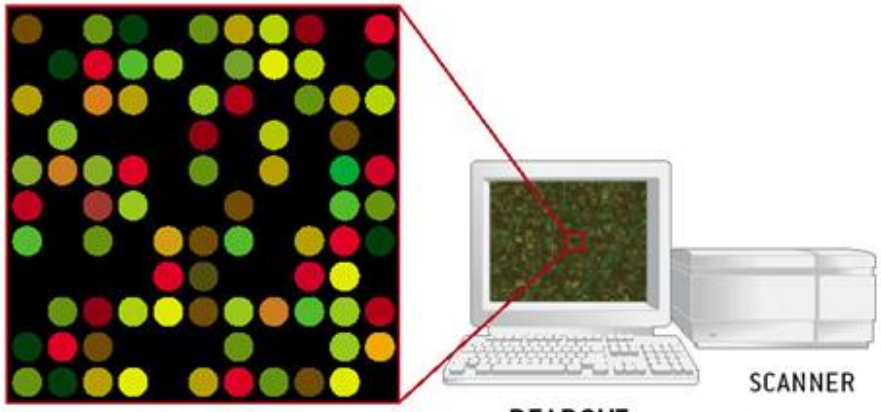
4 Apply the labeled cDNAs to the chip. Binding occurs when cDNA from a sample finds its complementary sequence of bases on the chip (*detail at right*). Such binding means that the gene represented by the chip DNA was active, or expressed, in the sample.



SCIENTIFIC AMERICAN EXPLAINS

Computational analysis.

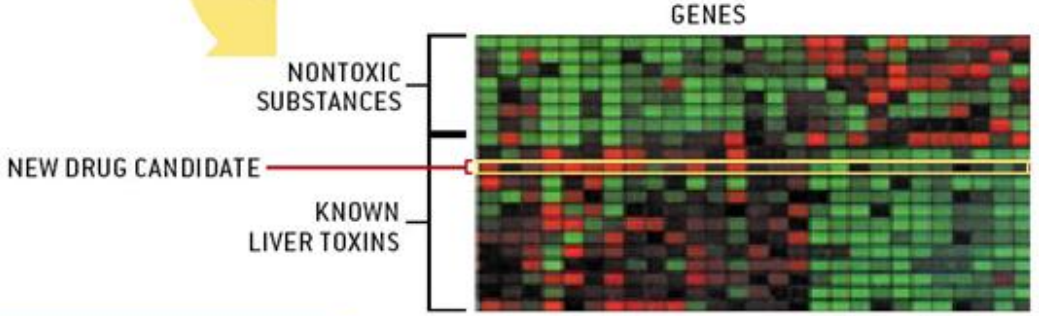
- GENE THAT STRONGLY INCREASED ACTIVITY IN TREATED CELLS
- GENE THAT STRONGLY DECREASED ACTIVITY IN TREATED CELLS
- GENE THAT WAS EQUALLY ACTIVE IN TREATED AND UNTREATED CELLS
- GENE THAT WAS INACTIVE IN BOTH GROUPS



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.

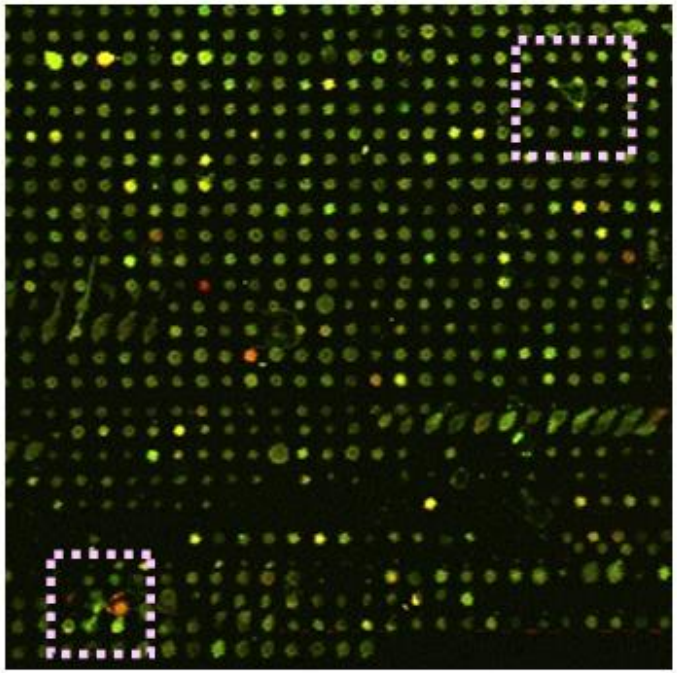
6 Determine whether any genes responded strongly to the drug in ways known to promote or reflect liver damage. Or compare the overall expression pattern produced by strong responders with the patterns produced when those genes react to known liver toxins (*right*). Close similarity would indicate that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

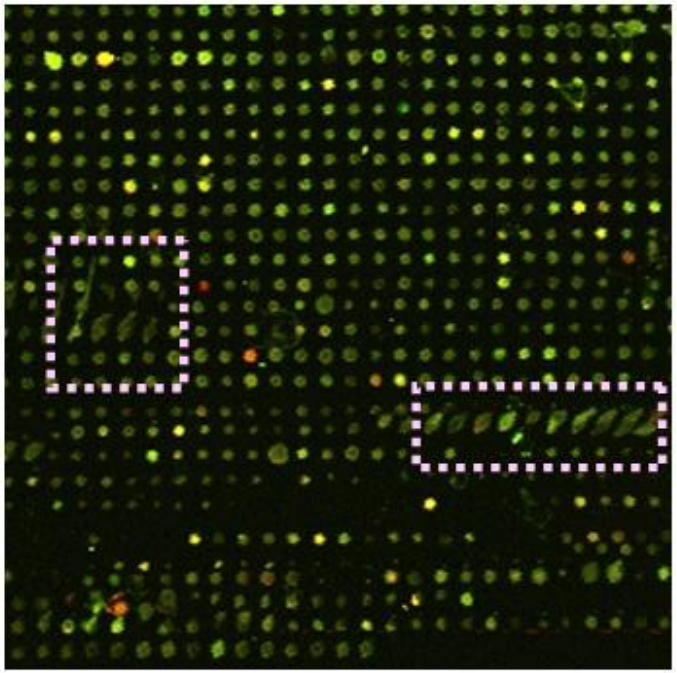
HYPOTHETICAL PROFILES OF GENE ACTIVITY IN CELLS TREATED WITH VARIOUS COMPOUNDS



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.

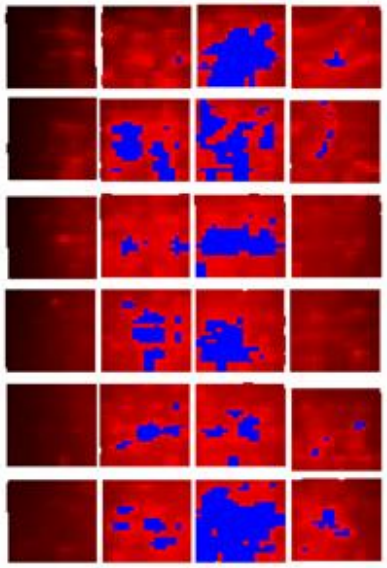
Background correction:

Raw data has large positional dependence.

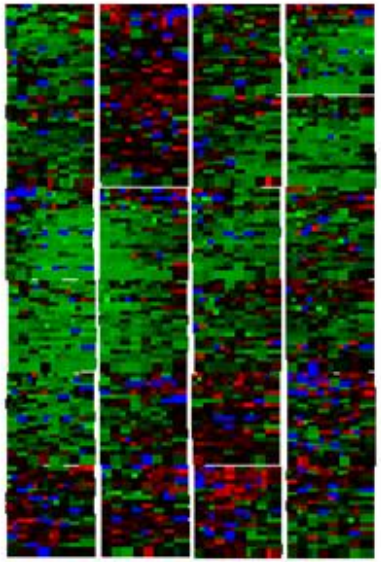
Remove artifactual smudges.

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

NOISY BACKGROUND



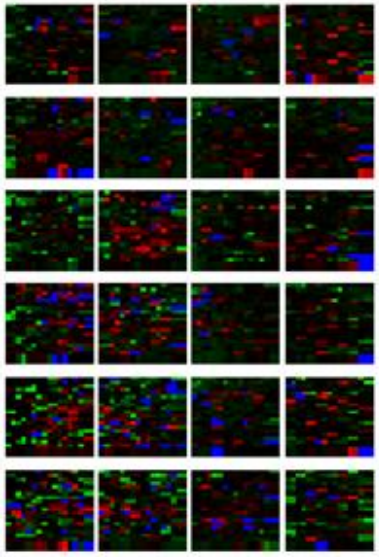
Before



Raw data has large positional dependence

After

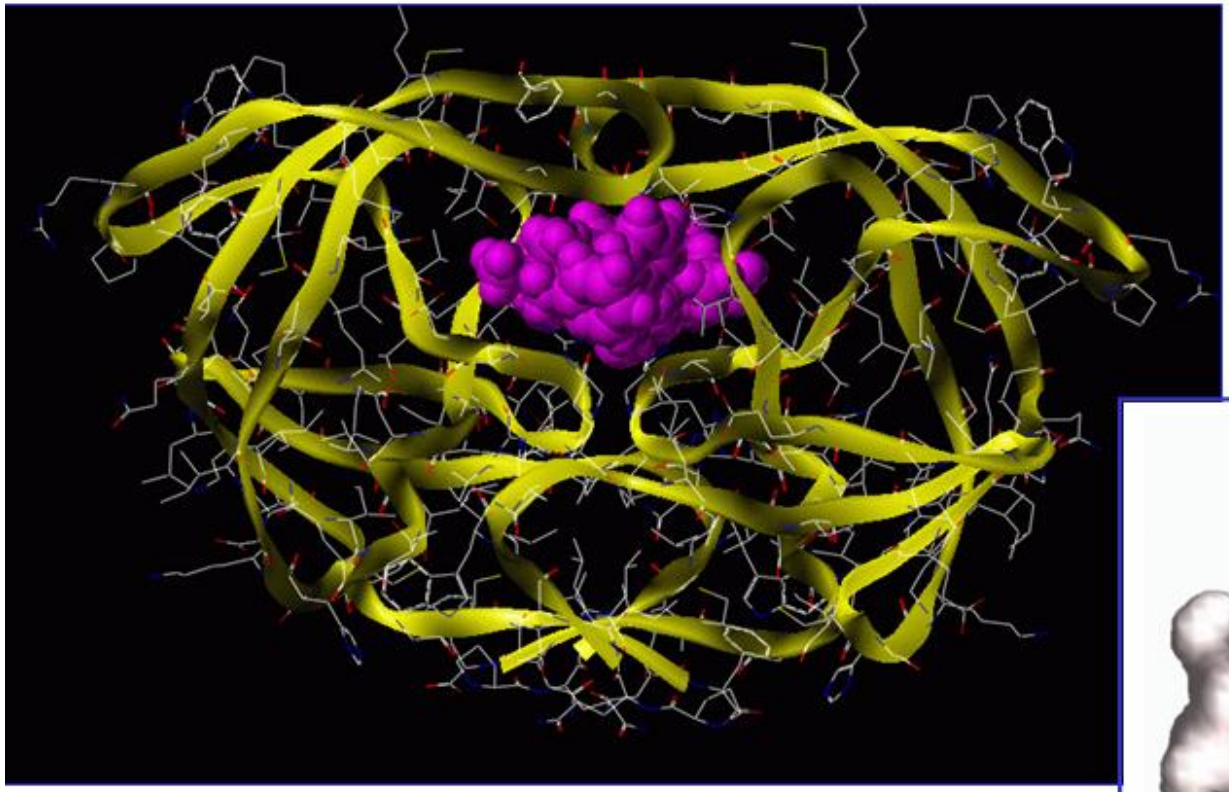
Remove artifactual smudges



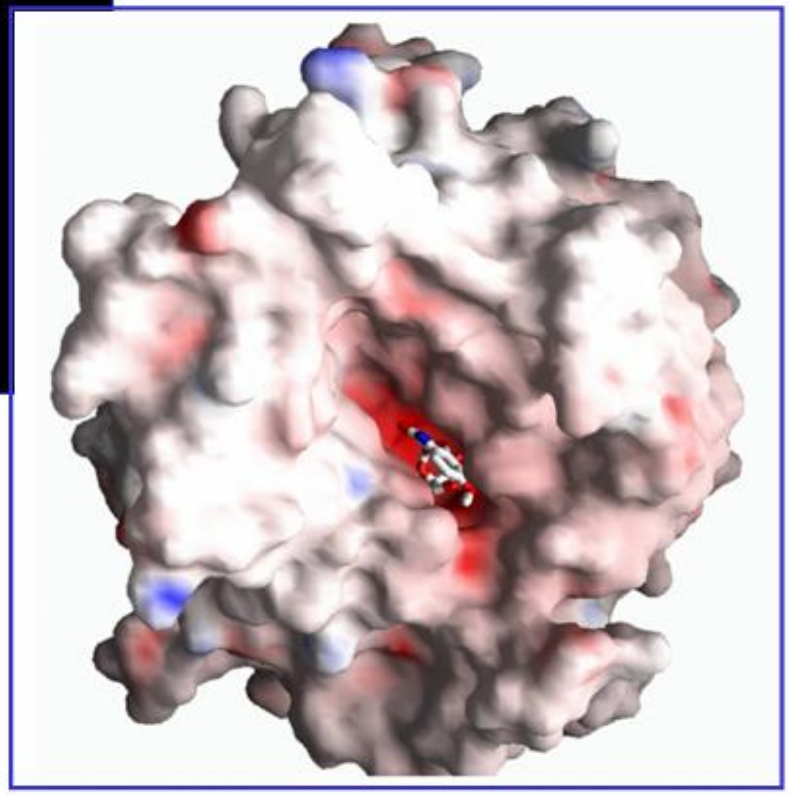
Positional normalization removes most of the artifact.

Discovering Drugs
Concept 7.5

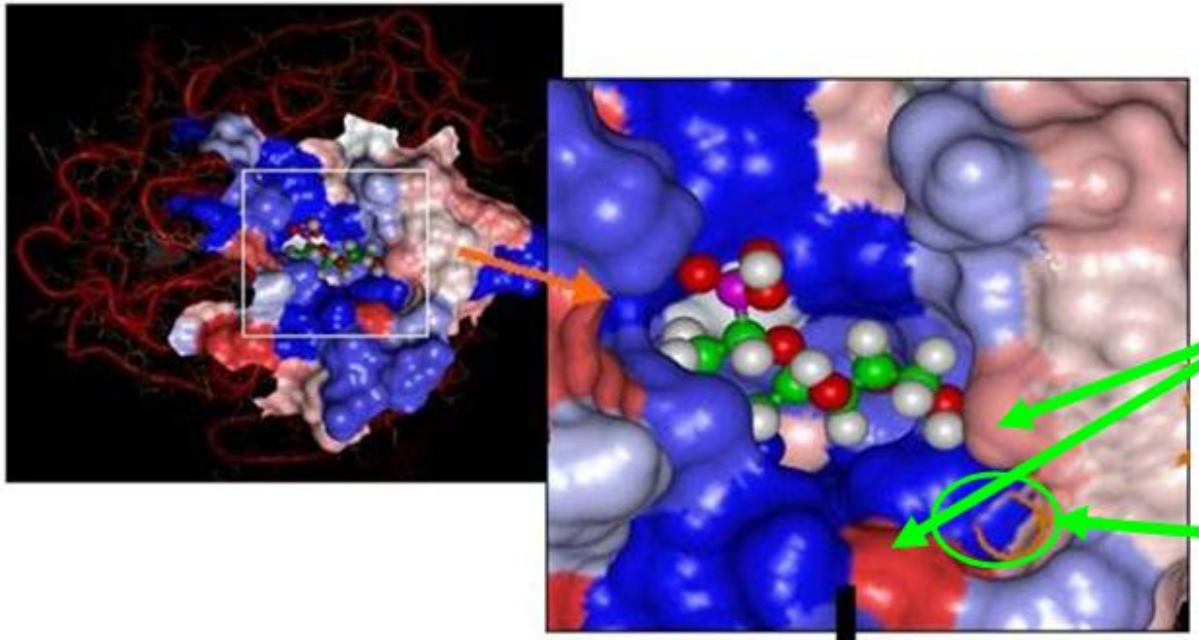
DRUG BINDING



Drugs bind to proteins by lock and key fit.

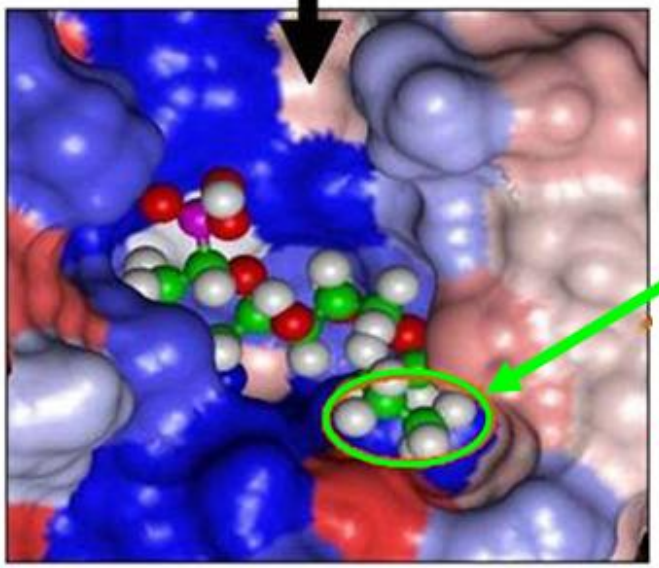


STRUCTURE BASED DRUG DESIGN



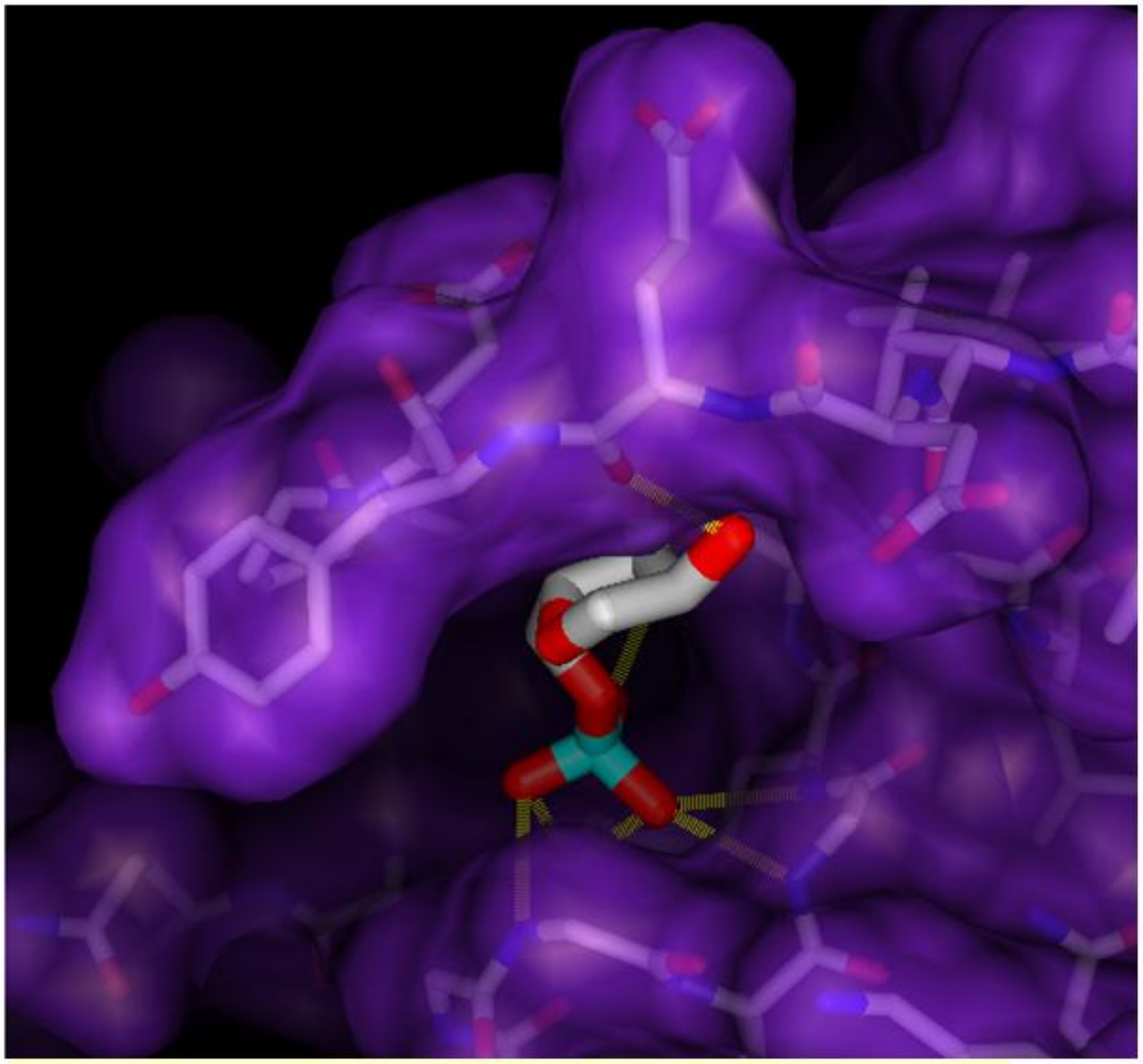
Hydrophobic residue.

Hydrophobic Pocket



Design drug to fit pocket

DRUG DOCKING



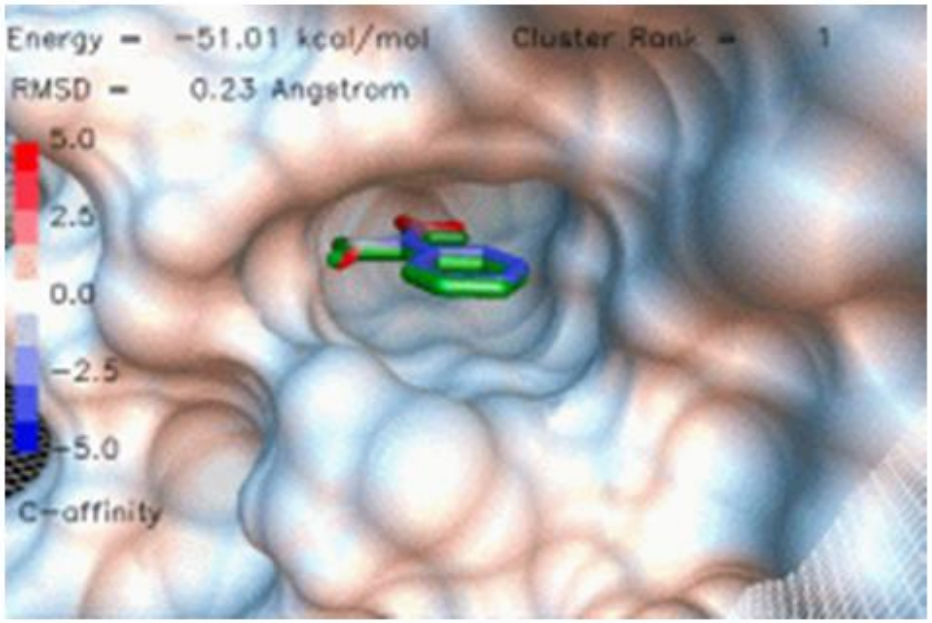
DOCK developed un 1981 by Kuntz and co-workers at UCSF.

Fit into active site by geometrical criteria

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

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

AUTODOCK

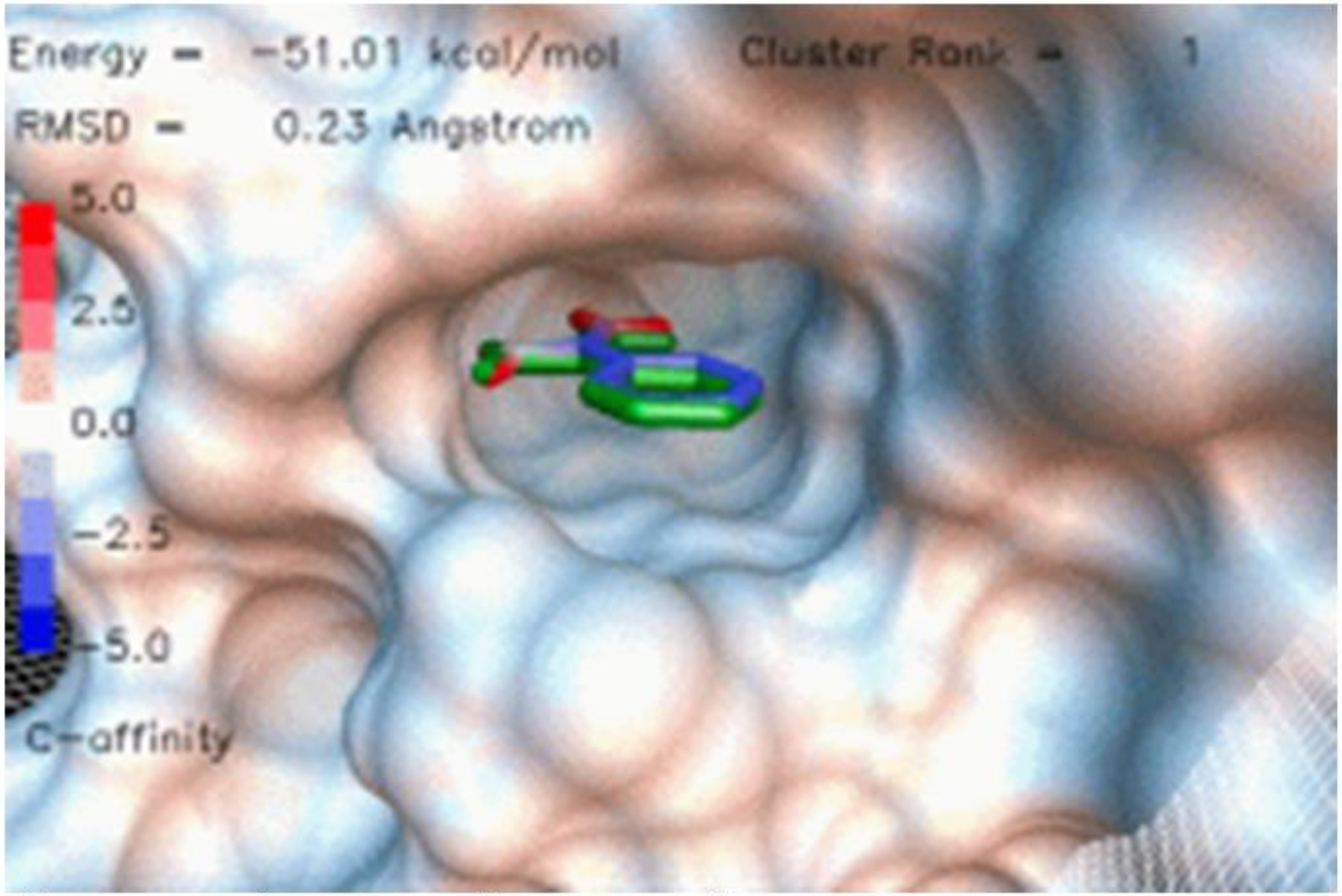


- Use simulated annealing to fit into active site with simple energy function.
- Olson group at Scripps in 1989.

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

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

AUTODOCK MOVIES

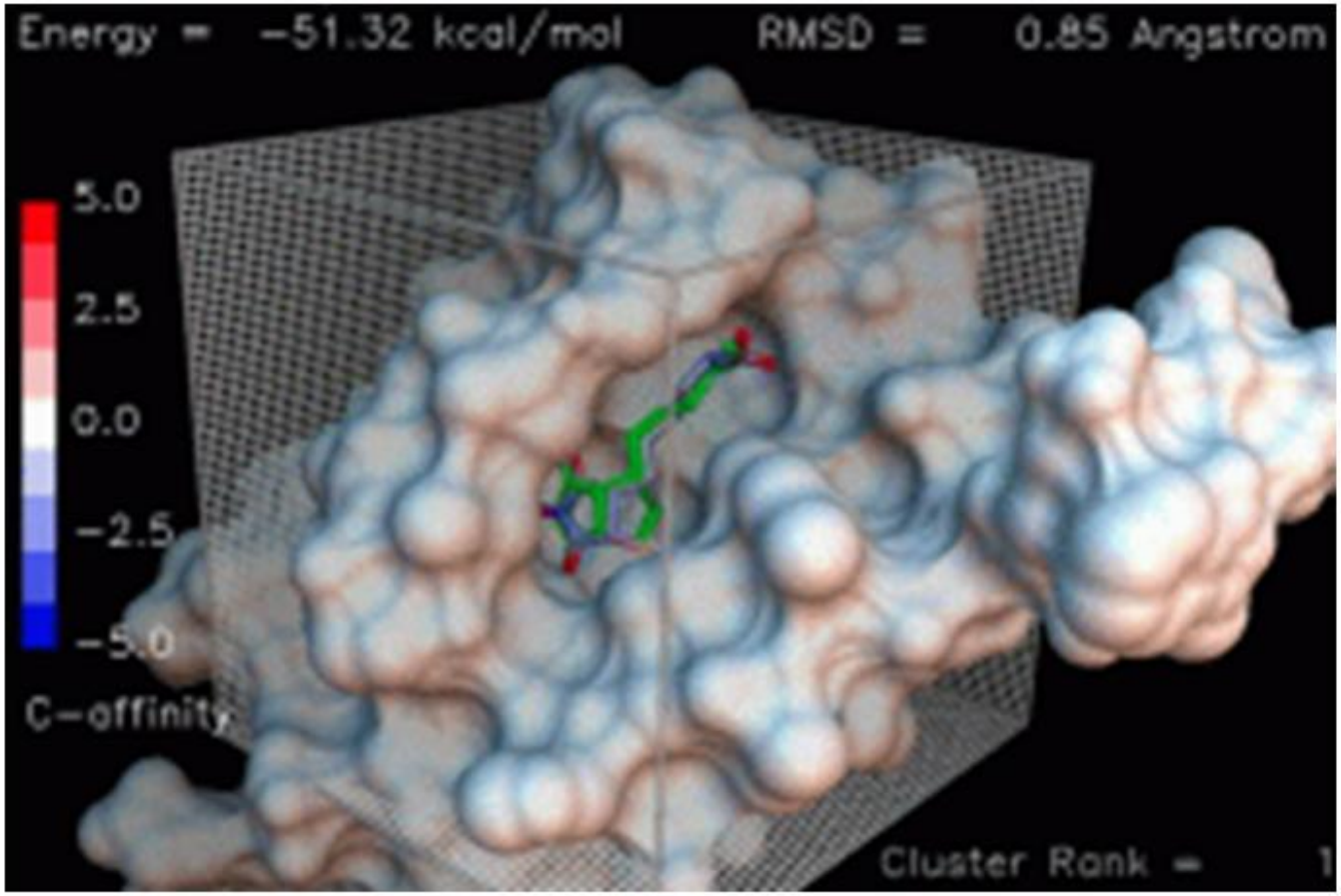


Benzamidine binding to Trypsin.

Gives energies in increasing order.

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

AUTODOCK MOVIES



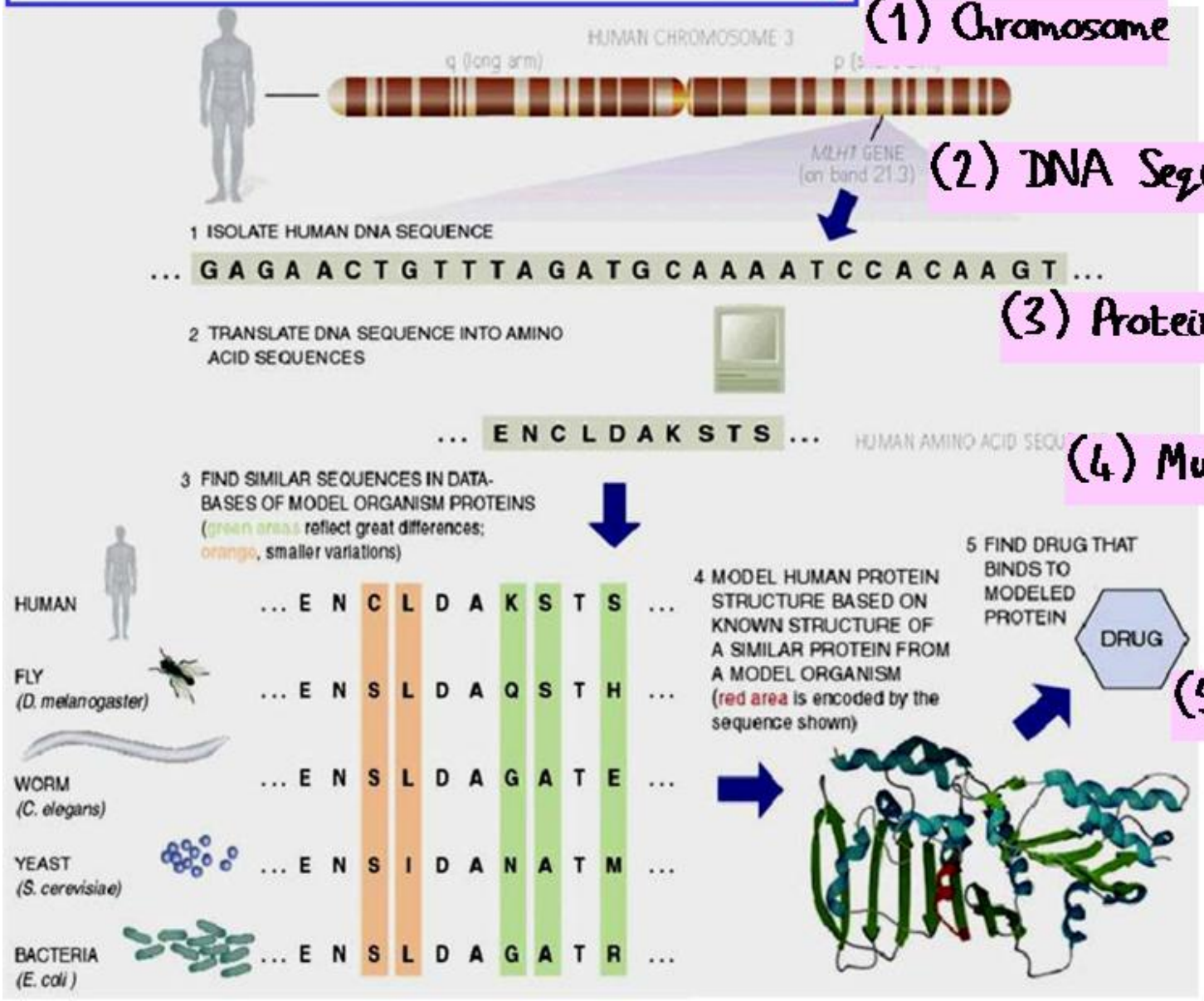
Biotin binding to Streptavidin.

Gives energies in increasing order.

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

DRUGS FROM SEQUENCE

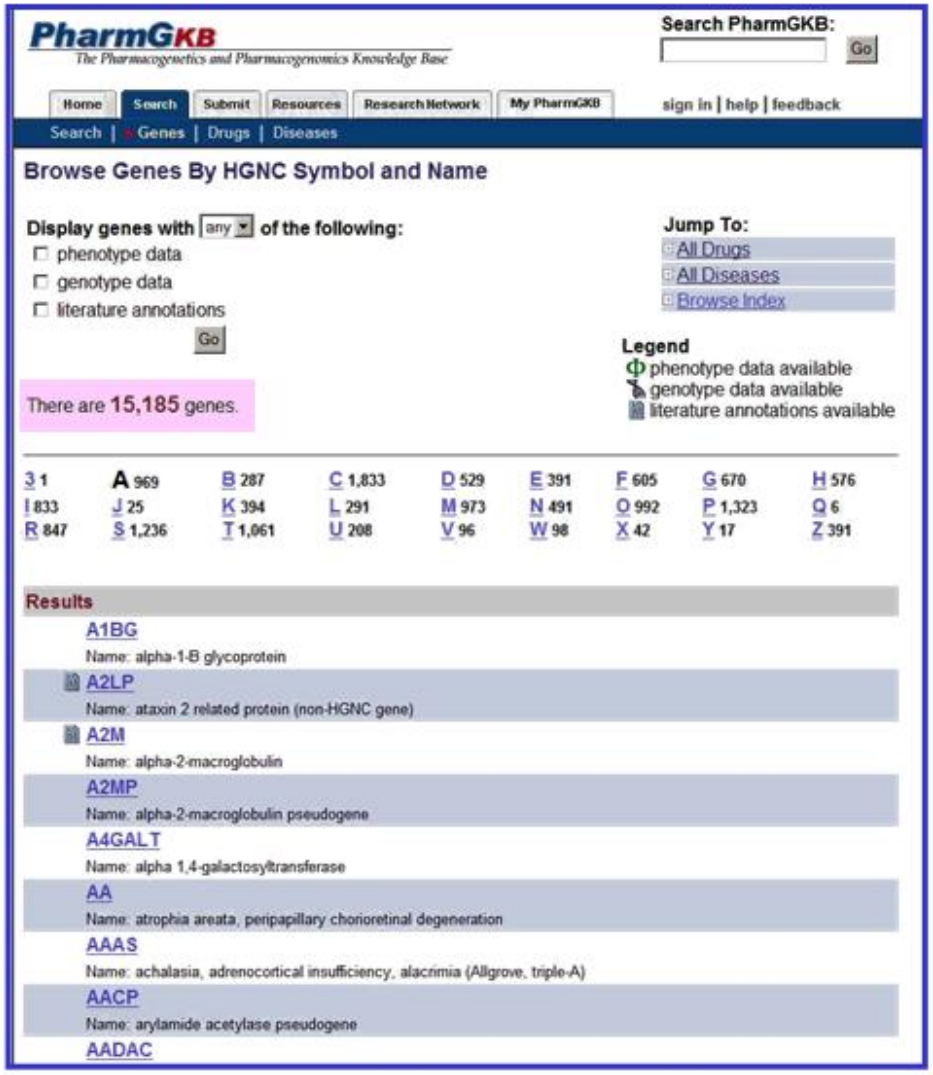
From Gene to Drug in six "easy" steps.



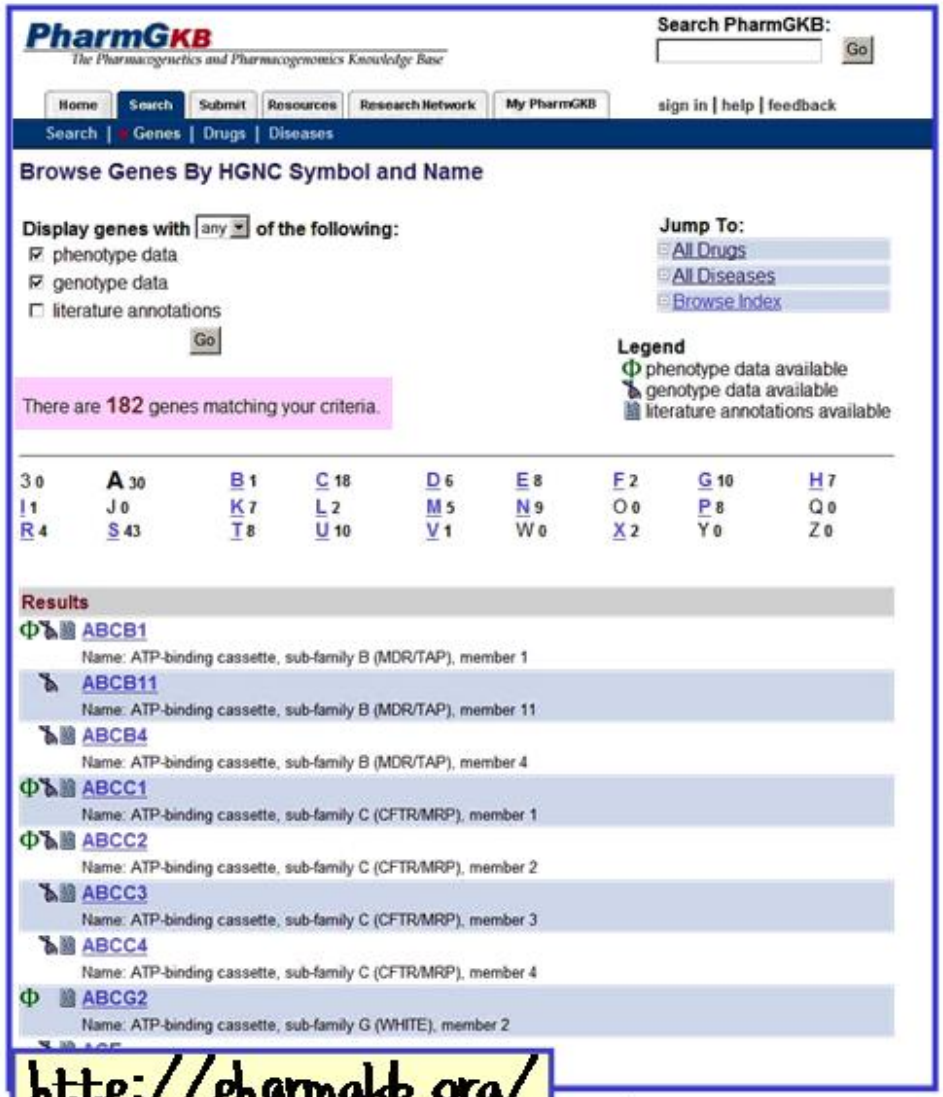
Diagnosing Disease
Concept 7.6

PHARMACOGENOMICS

Little data so far.



Lots of genes.



http://pharmgkb.org/

PROTEIN 2-D GELS

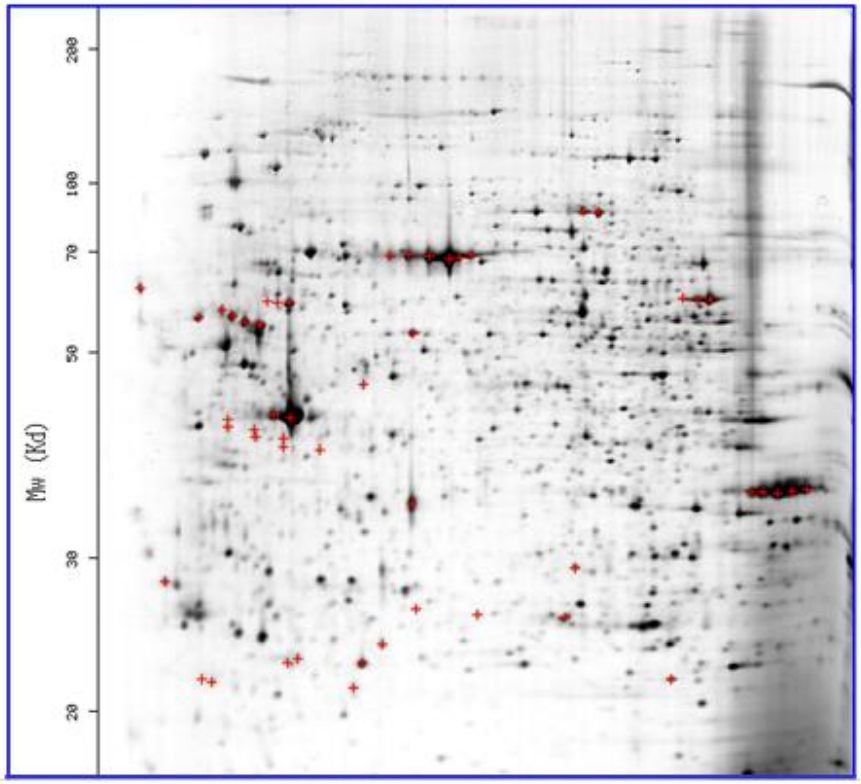
Map Selection: LYMPHOMA_HUMAN

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:

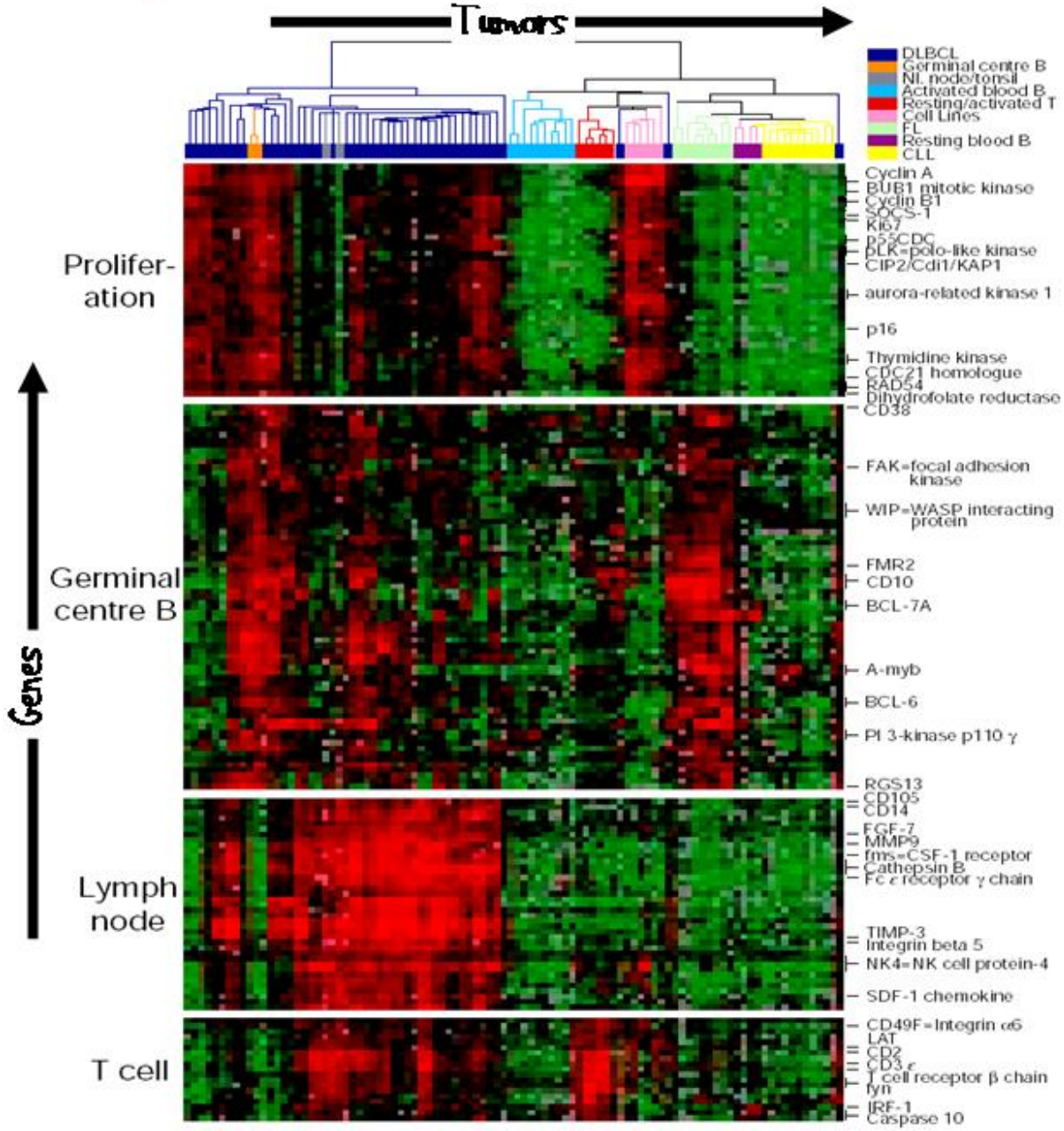
- LYMPHOMA_HUMAN large, spots highlighted
- LYMPHOMA_HUMAN small, spots not highlighted
- LYMPHOMA_HUMAN large, spots not highlighted

For high resolution gel image, see reference or download the tiff file.



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

EXPRESSION ARRAY DIAGNOSIS



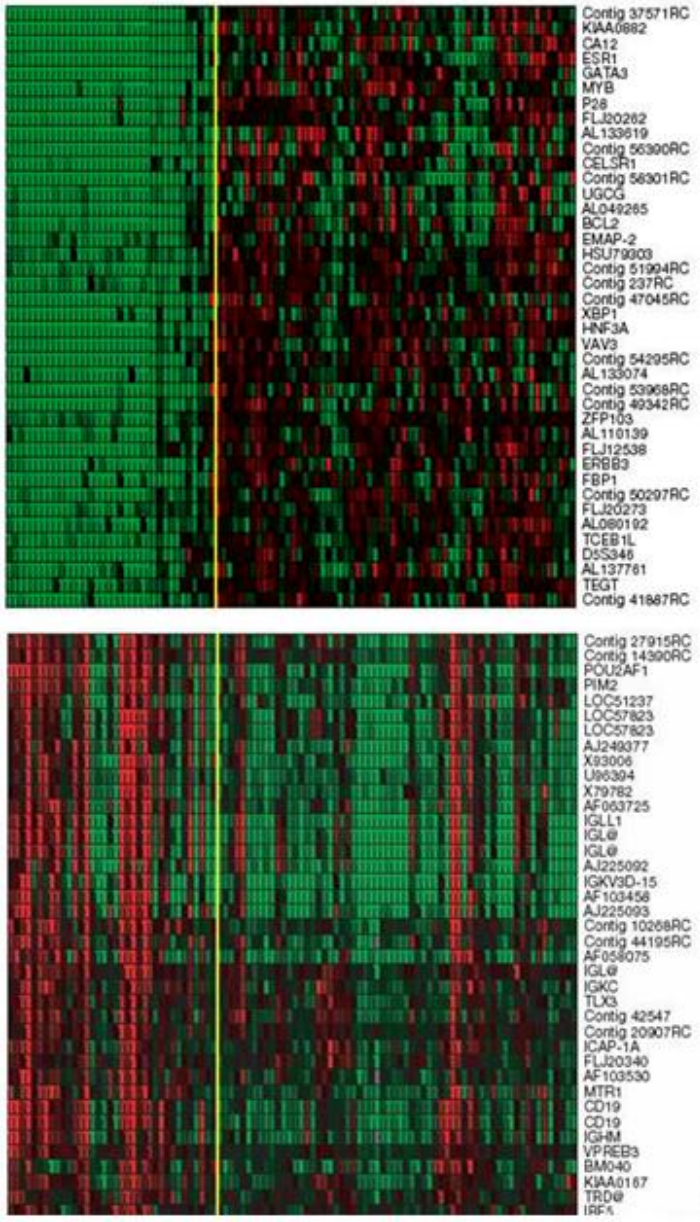
- Cluster genes by similarity of their expression pattern in the dfferent tumors
- Find that this clusters tumors.
- Proliferating cancers are separated from others by the genes that they express.

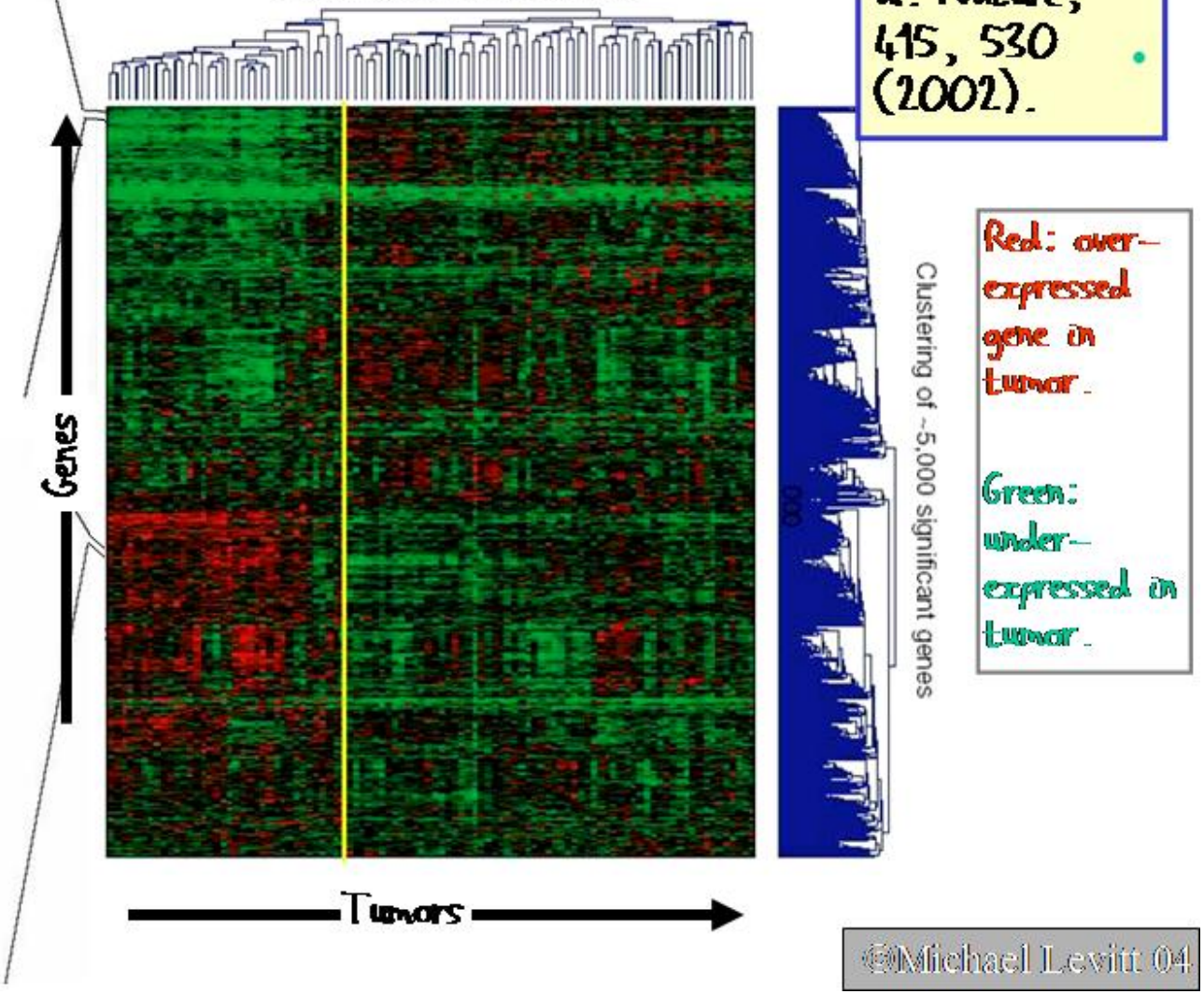
Red: over-expressed gene in tumor.
Green: under-expressed in tumor.

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

EXPRESSION ARRAY DIAGNOSIS

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





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

Red: over-expressed gene in tumor.

Green: under-expressed in tumor.