Hacking the Genome: Designer Proteins, Elite Organisms, and You

Russell Hanson

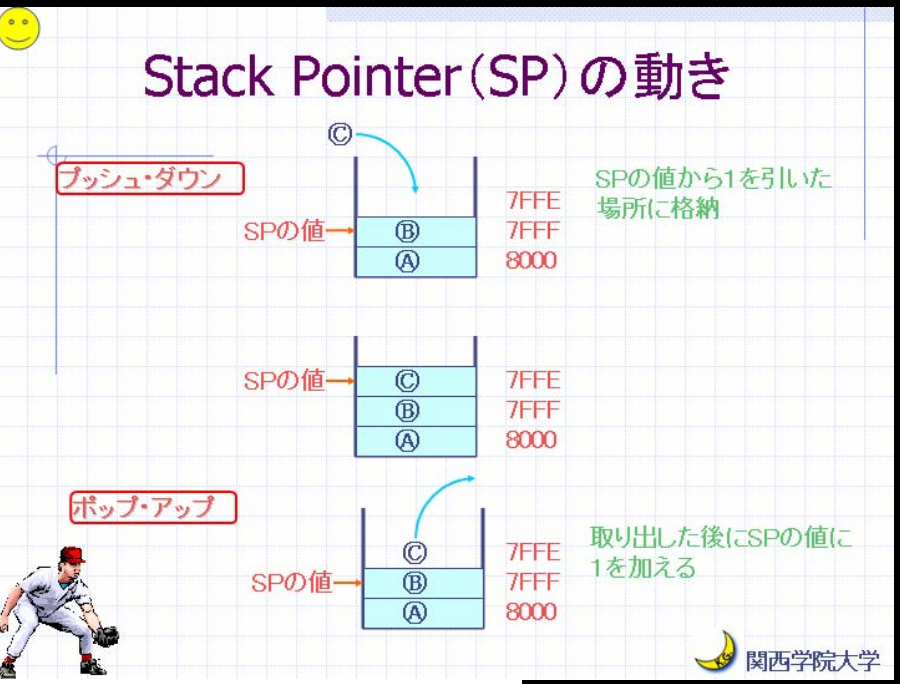
April 18, 2004

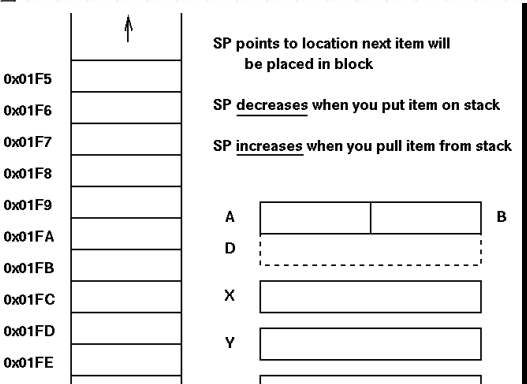
Outline

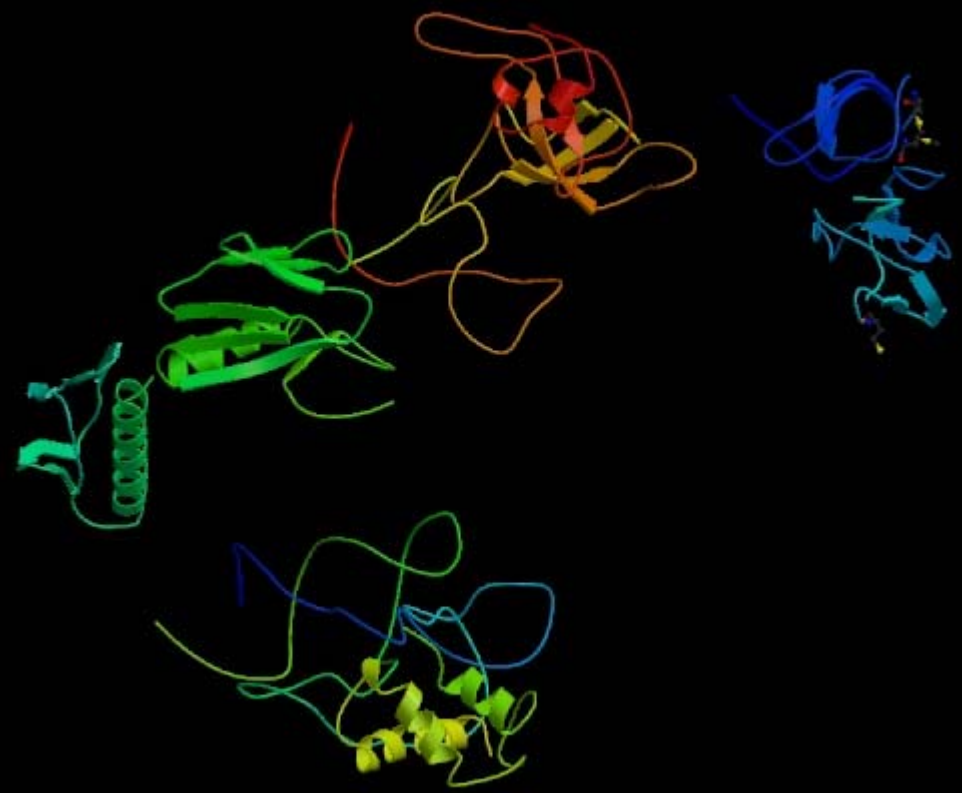
- Analogies – Why this talk?
- Engineering proteins
- Computer tools for genome analysis
- Conclusions

The Analogy

Instruction Pointer : Machine Code :: Ribosome : RNA







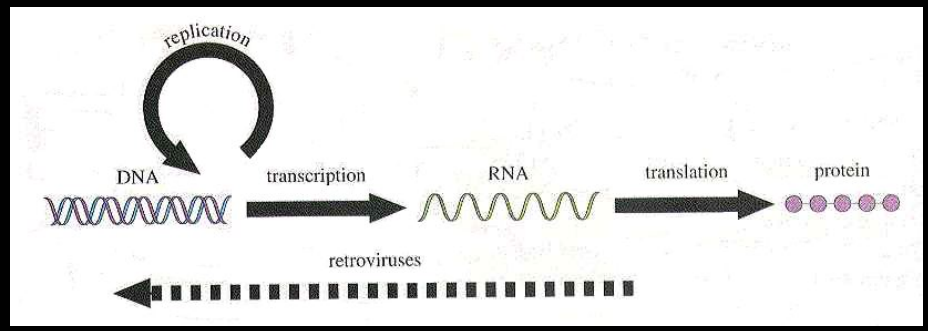
5 Å Map Of The Large Ribosomal Subunit

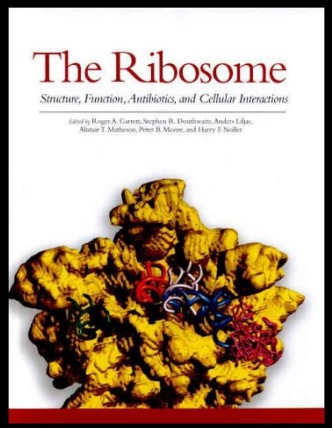
InterzOne III – Atlanta

The Analogies, cont.

Instruction Pointer : Machine Code ::
Ribosome : RNA

- The ribosome translates mRNA to polypeptides (transcription -> RNA-processing of pre-mRNA ->mRNA translation)





R. Garrett *et al.* The Ribosome: Structure, Function, Antibiotics, and Cellular Interactions (2000)

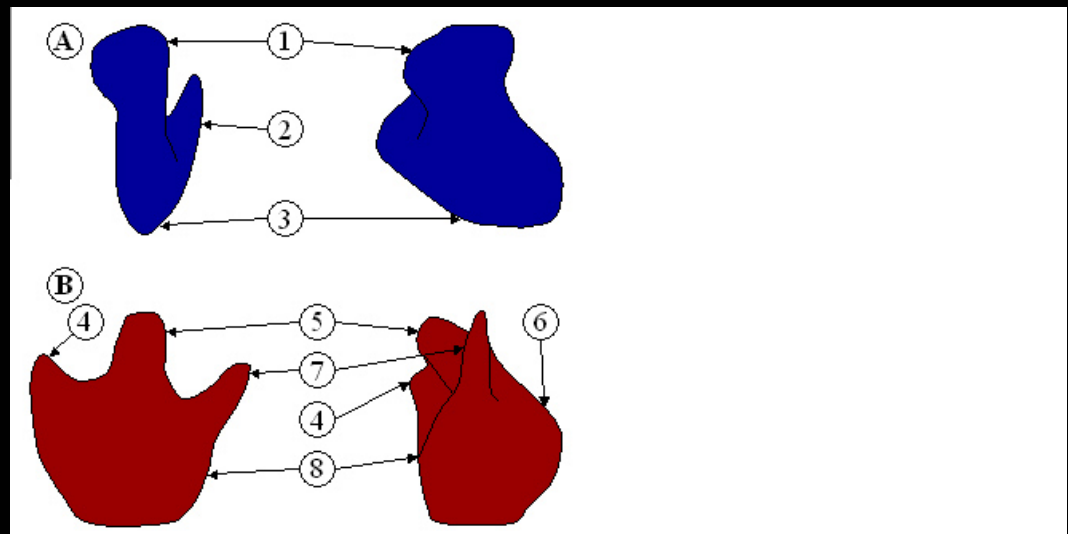
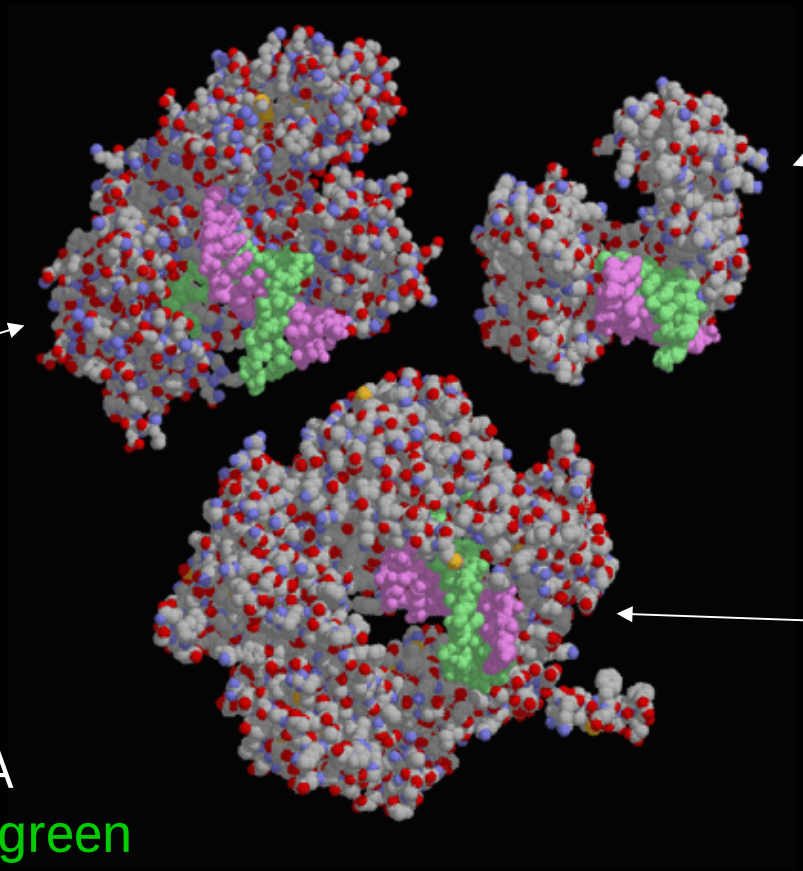


Figure 1 : The subunits of a ribosome. Side and front view.
(A) Small subunit. (B) Large subunit. (1) Head. (2) Platform. (3) Base. (4) Ridge. (5) Central protuberance. (6) Back. (7) Stalk. (8) Front.

More Analogies

I) Canonical shell commands: cp, mv, cc, ar, ln, ld, gprof, ...

II) Biological functional elements: DNA Polymerase, ATP/GTP Powered Pumps, Ribosome, Signal transduction pathways, measure macroscopic gene expression, ...



H. Sapiens

PDB: 1zqa

E. Coli

PDB: 1kln

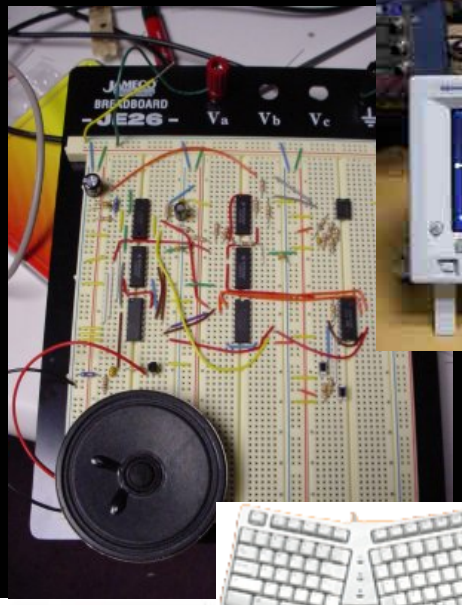
Viral

PDB: 1clq

DNA Polymerase

Small piece of DNA bound is purple & green

hACKER Lab vs. Bio Lab



Machines

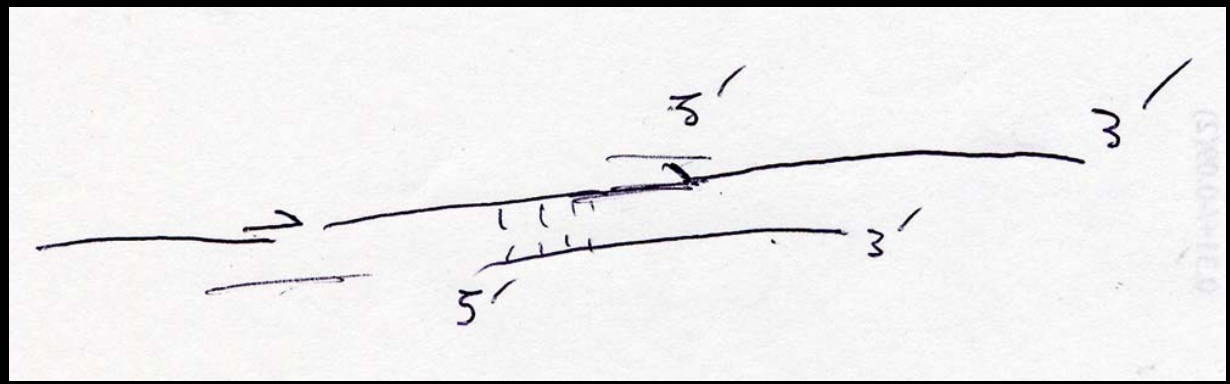
- DNA sequence synthesis
- Online can buy for $.50/bp, up to 45 nucleotide length fragment.
- Buy your own peptide/nucleotide synthesizer for $500-$25K USD.

DNA Synthesis - Beckman Oligo 1000

Peptide Synthesis - Applied Biosystems 431A

PCR lets you assemble pieces *ad infinitum*

- Sketch:



Applied BioSystems Real-Time PCR machine ($25K-$45K)

Engineering

- Engineer a protein
- Engineer an organism

.... Why?

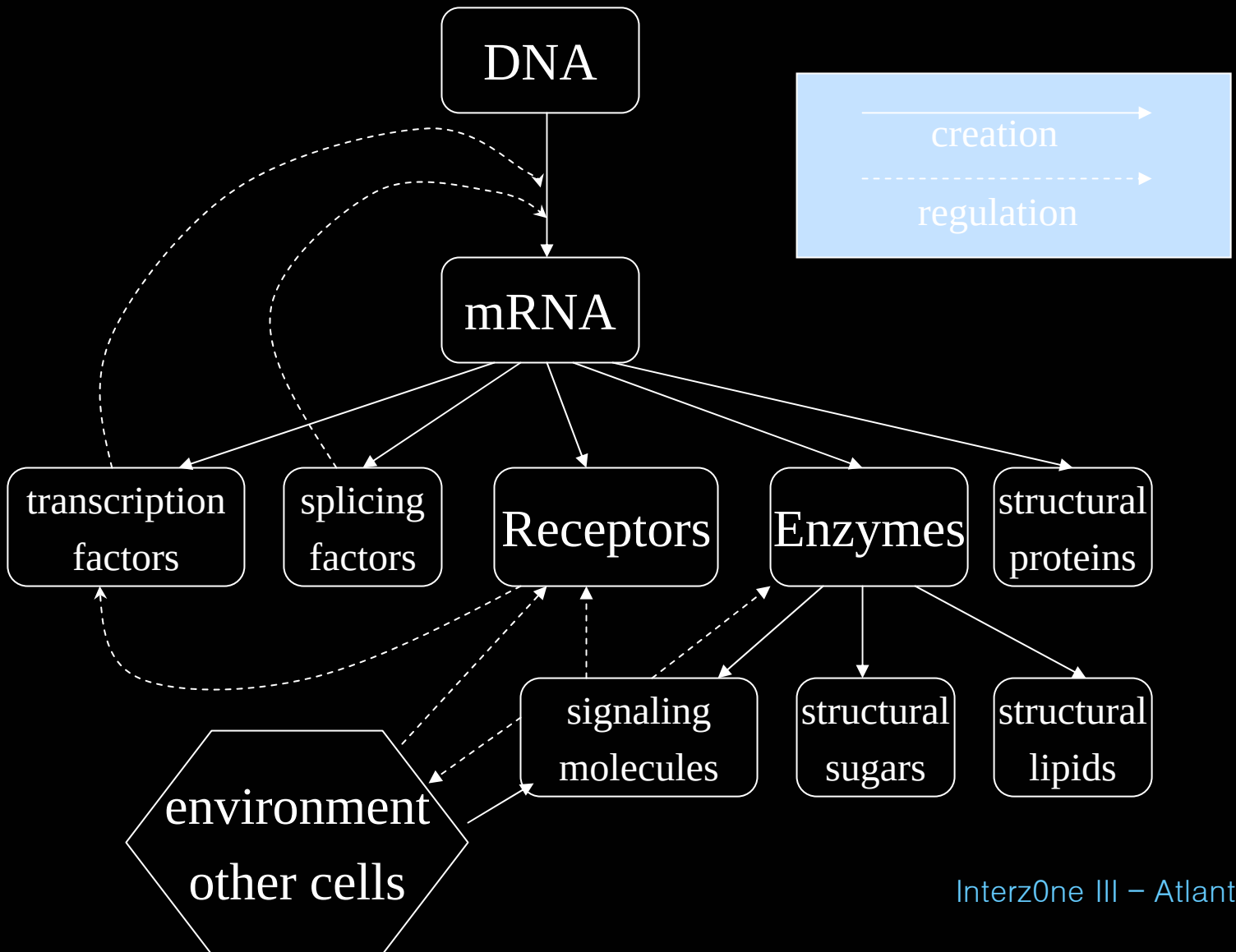
"There is at present no understanding of this hacker mindset, the joy in engineering for its own sake, in the biological community."

-Roger Brent (Cell 2000)

Oh, *engineered* organisms

- Corn
- Tomatoes
- Citrus fruit
- (...)
- And our friend, the fruit fly, *Drosophila Melanogaster*
- Celera, Inc. released information on *genomic*-scale engineering, not available at press time

Primary Flows of Information and Substance in a Cell



Review: Protein... hunh?

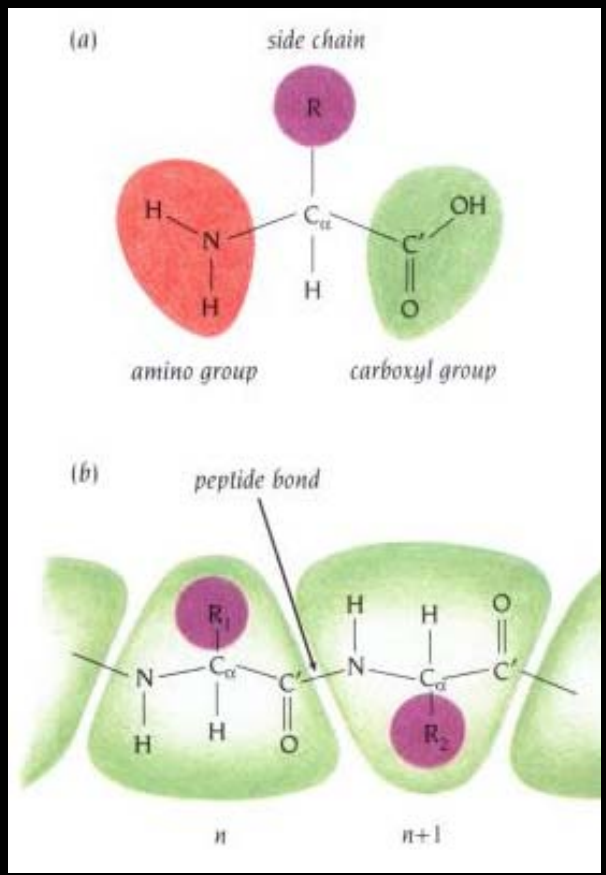


Figure 1.2 Proteins are built up by amino acids that are linked by peptide bonds to form a polypeptide chain. (a) Schematic diagram of an amino acid, illustrating the nomenclature used in this book. A central carbon atom (C_α) is attached to an amino group (NH_2), a carboxyl group (COOH), a hydrogen atom (H), and a side chain (R). (b) In a polypeptide chain the carboxyl group of amino acid n has formed a peptide bond, C–N, to the amino group of amino acid $n + 1$. One water molecule is eliminated in this process. The repeating units, which are called residues, are divided into main-chain atoms and side chains. The main-chain part, which is identical in all residues, contains a central C_α atom attached to an NH group, a C′=O group, and an H atom. The side chain R, which is different for different residues, is bound to the C_α atom.

Why engineer proteins?

- 1) Engineered macromolecules could have experimental use as experimental tools, or for development and production of therapeutics
- 2) During the process of said engineering, new techniques are developed which expand options available to research community as whole
- 3) By approaching macromolecule as engineer, better understanding of how native molecules function

(Doyle, Chem & Bio, 1998)

Is this how a “hacker” approaches a problem?

- 1) determine what are elemental tools/components, learn to work with them, develop something new
- 2) engineering in the large
- 3) note however the physics/chemistry of proteins, the Levinthal paradox, and the amount of effort spent on protein folding, i.e. “more time to hack”

Levinthal Paradox (1968):

given a peptide group 3 possible conformations of

bond angles φ and ψ, in allowable regions

given a protein of 150 amino acids

$= 3^{150}$ possible structures $\sim= 10^{68}$

time of bond rotation 10^{-12}s

$10^{68} * 10^{-12}$s $= 10^{56}$sec$=10^{48}$ years

Life on earth $3.8 * 10^{9}$ years

Real folding times are 0.1 – 1000 sec

Methods for *de novo* protein synthesis

Two methods:

TASP: Template-assembled synthetic proteins

RAFT: Regioselectively addressable functionalized templates

"Small proteins or protein domains that are structurally stable and functionally active are especially attractive as models to study protein folding and as starting compounds for drug design, but to select them is a difficult task.

…

Advances in protein design and engineering, synthesis strategies, and analytical and conformational analysis techniques allowed for the successful realization of a number of folding motifs with tailored functional properties."

(Tuchscherer, Biopolymers, 1998)

Adding functional motifs to stable structures

(Tuchscherer, Biopolymers, 1998)

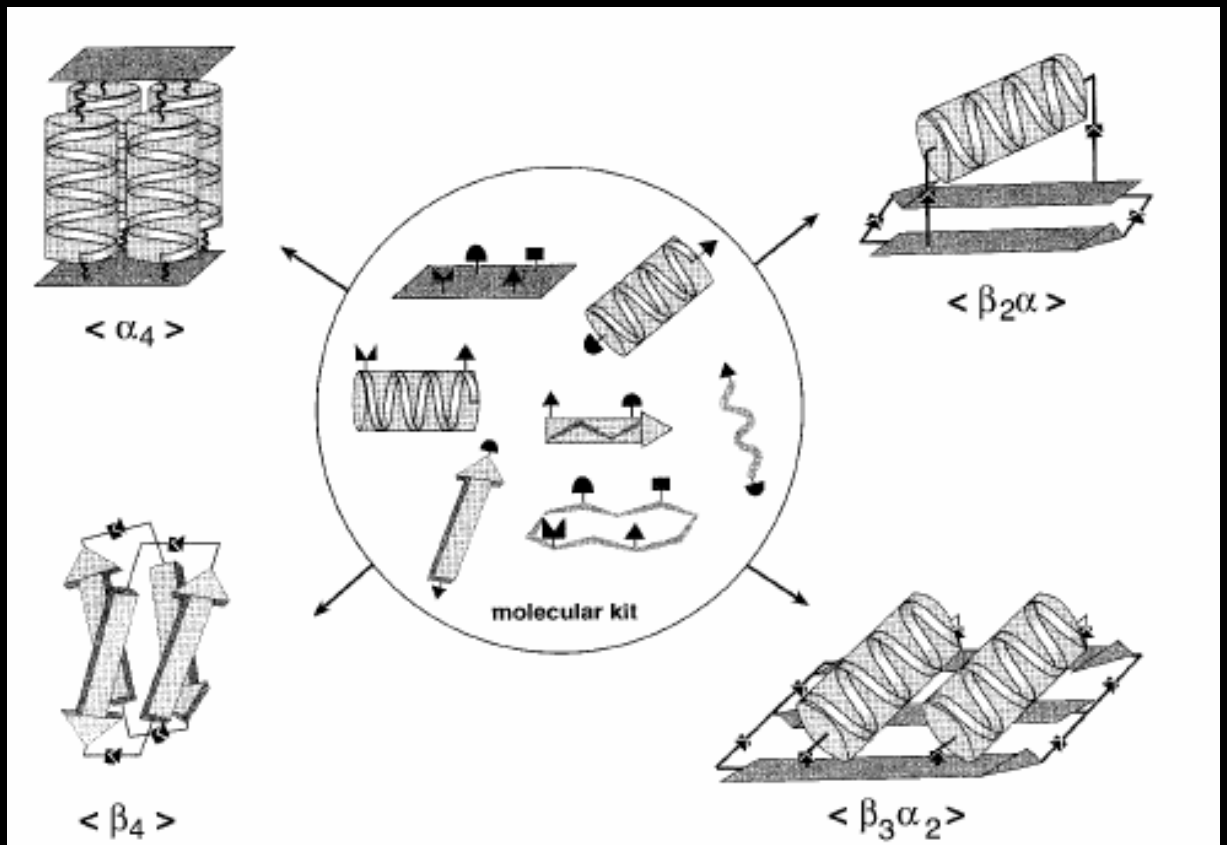


FIGURE 2 Locked-in tertiary folds as extension of the TASP concept: by applying the principles of a molecular kit, individual secondary structure elements such as helices, β-sheets, turns, and loops are covalently attached via both chain ends to appropriately functionalized templates. The resulting multibridged molecules, e.g., locked-in 4-helix $\langle\alpha 4\rangle$ and β-sheet $\langle\beta_4\rangle$ bundles, $\beta\beta\alpha$- $\langle\beta_2\alpha\rangle$ or more complex arrangements, e.g., $\langle\beta_3\alpha_2\rangle$, are molecules with a built-in pathway for folding.[22,36]

Ligand Binding – protein flexibility

"In this study, we set out to elucidate the cause for the discrepancy in affinity of a range of serine proteinase inhibitors for trypsin variants designed to be structurally equivalent to factor Xa." (Rauh, J. Mol. Biol., 2004)

Def: Ligand

Any molecule that binds specifically to a receptor site of another molecule; proteins embedded in the membrane exposed to extracellular fluid.

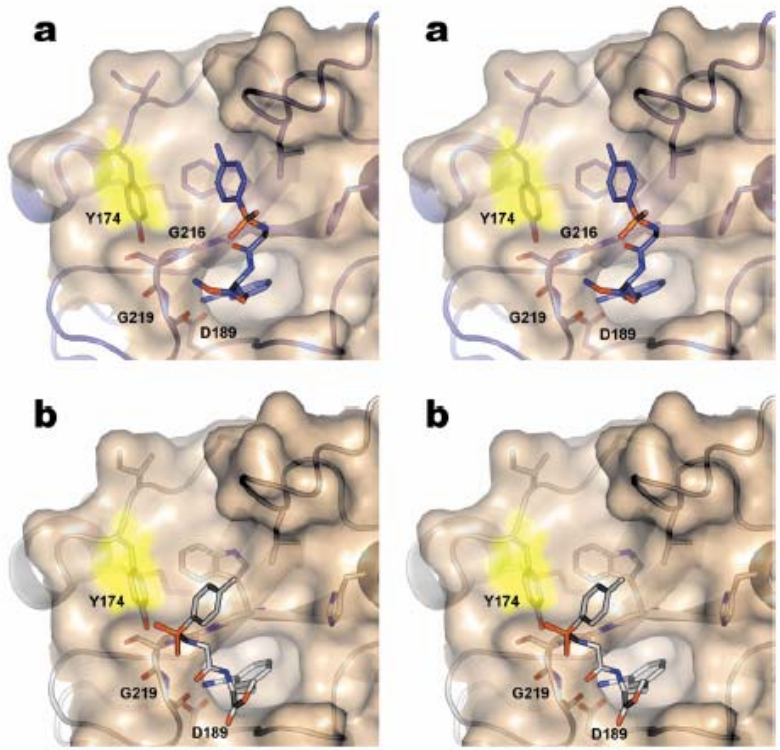


Figure 2. Stereo view showing the alternative binding modes adopted by inhibitor (4) in the two structures (a) X(SSYI)bT.A4 and (b) X(SSYI)bT.B4. In a, the glycine spacer hydrogen bonds to Gly216; the tosyl group of the inhibitor occupies the S3/S4 site. In b, the glycine spacer hydrogen bonds to Gly219; the tosyl group points away from the enzyme, making contacts with a symmetry-related molecule in the crystal (not shown).

One way to test for ligand binding

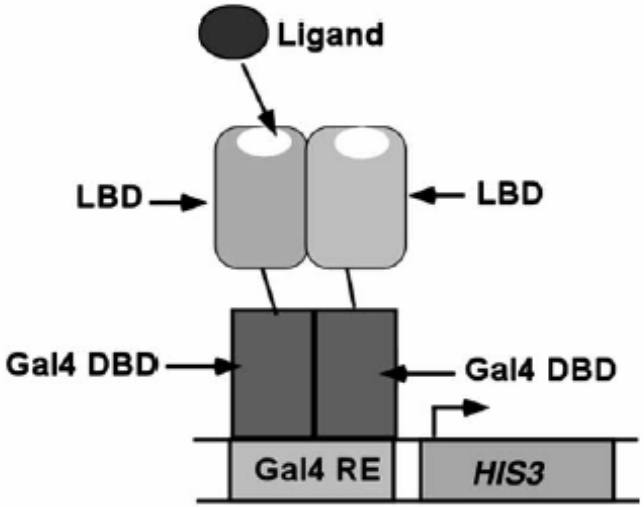
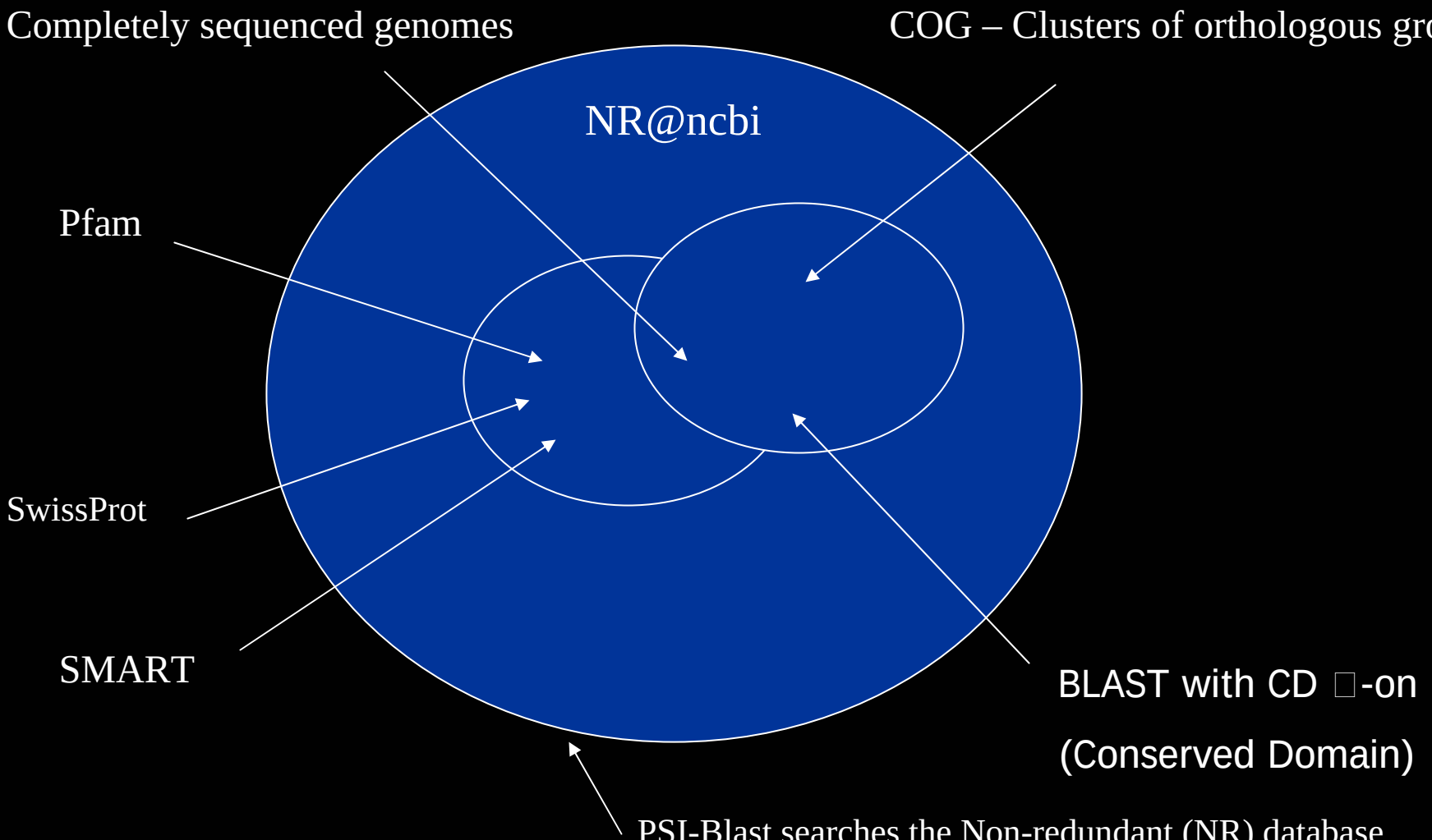


Fig. 1. Genetic selection using *S. cerevisiae* strain PJ69-4A. In this system the nuclear receptor's ligand binding domain (LBD) is fused to the Gal4 DNA binding domain (Gal4 DBD). The fusion protein binds to the Gal4 response element controlling the expression of the *HIS3* gene and if the *HIS3* gene is expressed, yeast cells are able to grow on media minus histidine. By transforming the expression plasmids coding for the nuclear receptors into the yeast strain and plating them onto plates minus histidine but containing the appropriate ligand, the nuclear receptor activates transcription of *HIS3* gene. In a process analogous to classical genetic complementation, the small molecule complements the histidine auxotroph, allowing the yeast to survive through a process termed "chemical complementation."

(Doyle, Biochemical and Biophysical Research Comm., 2003)

Bioinformatics Databases



How to Access the Human Genome (and other sequenced genomes)

- ftp://blah.blah.blah.blah

Index of ftp://ftp.ncbi.nih.gov/genbank/genomes

Up to higher level directory

A thaliana		10/17/2003 0:00:00
Anopheles gambiae		5/7/2002 0:00:00
Bacteria		4/7/2004 18:28:00
C elegans		6/14/2002 0:00:00
D melanogaster		10/19/2000 0:00:00
H sapiens		4/15/2004 0:23:00
Leptospira interrogans serovar Copenhageni		3/22/2004 17:40:00
MITOCHONDRIA		11/2/1999 0:00:00
M musculus		5/12/2002 0:00:00
P falciparum		5/11/1999 0:00:00
Plasmodium falciparum		10/11/2002 0:00:00
README		
README OLD		
R norvegicus		
S cerevisiae		

```
hs_phs0.fna.gz Survey sequence (appr
hs_phs1.fna.gz Unordered contigs (ea
hs_phs2.fna.gz Ordered contigs (each
hs_phs3.fna.gz Finished sequence
```

Index of ftp://ftp.ncbi.nih.gov/genbank/genomes/H_sapiens

Up to higher level directory

README	2 KB	10/17/2003	0:00:00
hs phase0.fna.gz	94408 KB	4/14/2004	21:36:00
hs phase1.fna.gz	606370 KB	4/14/2004	22:35:00
hs phase2.fna.gz	48487 KB	4/14/2004	22:39:00
hs phase3.fna.gz	1090520 KB	4/15/2004	0:23:00

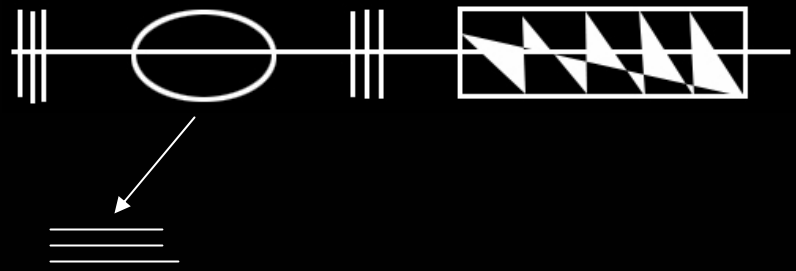
How to analyze a genome, or subsequence (p1)

- 1st Step: a) Working with unknown protein sequence; BlastP with CD on; you're finding similarity to other proteins, similarity of entire AA sequence

  b) COGnitor, precomputed BLASTs; metabolic pathways annotated; COGnitor more sensitive since 1) found similarities in BLAST, pulled them out 2) works on domain level
- 2nd Step: SEG (filtering of low-complexity segments); run COILS find α-helices; run SignalP find signal peptides; intrinsic properties of SMART, DAS
- 3rd Step: run PSI-BLAST to convergence; Pfam picks up 60% of known homologs (genes with common ancestor); started with few genomes

How to analyze a genome, or subsequence (p2)

- 4th Step: take result from PSI-BLAST; run Multiple Alignment on that; run Consensus (http://www.accelrys.com/insight/consensus.html) to find conserved regions



- 5th Step: Predict secondary structure: http://www.compbio.dundee.ac.uk/~www-jpred/
  - Prediction method: "Jnet; two fully connected, 3 layer, neural networks, the first with a sliding window of 17 residues predicting the propensity of coil, helix or sheet at each position in a sequence. The second network receives this output and uses a sliding window of 19 residues to further refine the prediction at each position."
  - Determine if protein of *unknown function*; make inferences based on structure prediction

PSI-BLAST

http://www.ncbi.nlm.nih.gov/BLAST/

- A normal BLASTP (protein-protein) run is performed.
- A position-dependent matrix is built using the most significant matches to the database.
- The search is rerun using this profile.
- The cycle may be repeated until convergence.
- The result is a ‘matrix’ tailored to the query.

Evolutionary Genomics

- From a phylogenetic tree can infer inheritance of proteins, and thereby organisms (conserved vs. non-conserved domains, etc).

Definitions:

homologs: if two genes/proteins share a common evolutionary history (not nec. same function)

analogs: proteins that are not homologs, but perform similar function

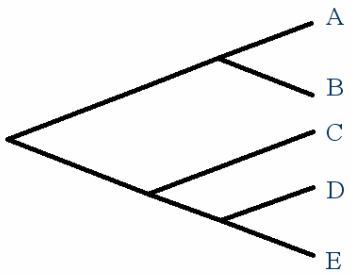
paralogs: products of gene duplication

orthologs: genes that are derived vertically, no guarantee that perform same function

Three types of trees

The Three Types of Trees

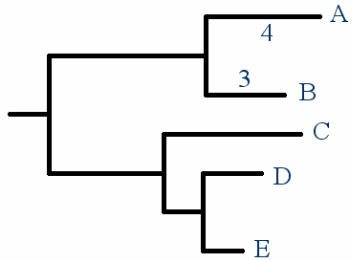
- Cladogram
  - Relative recency of common ancestry
  - No measurement of time or change



((A,B),(C,(D,E)))

The Three Types of Trees

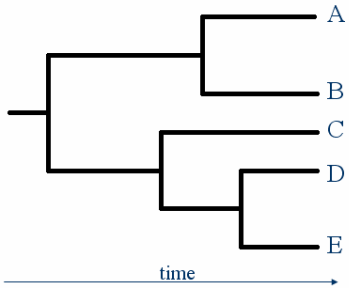
- Additive tree (Phylogram)
  - Relative recency of common ancestry
  - Branch length contains additional information, typically related to the amount of change between sequences



((A,B),(C,(D,E)))

The Three Types of Trees

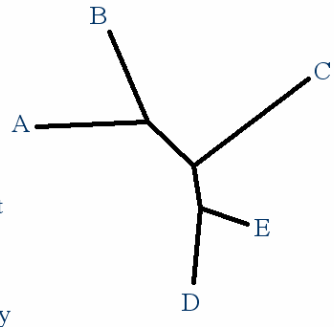
- Ultrametric tree (Dendrogram)
  - Relative recency of common ancestry
  - Depicts evolutionary time directly as years or indirectly as amount of sequence divergence via molecular clock



((A,B),(C,(D,E)))

Rooted vs Unrooted Trees

- Cladograms and phylograms can be either rooted or unrooted
  - Cannot define ancestors and descendants in the same manner
  - Can still distinguish clusters
    - Particularly useful in looking at different functions of related proteins.
  - Root of a tree is not necessarily assigned correctly by the program.

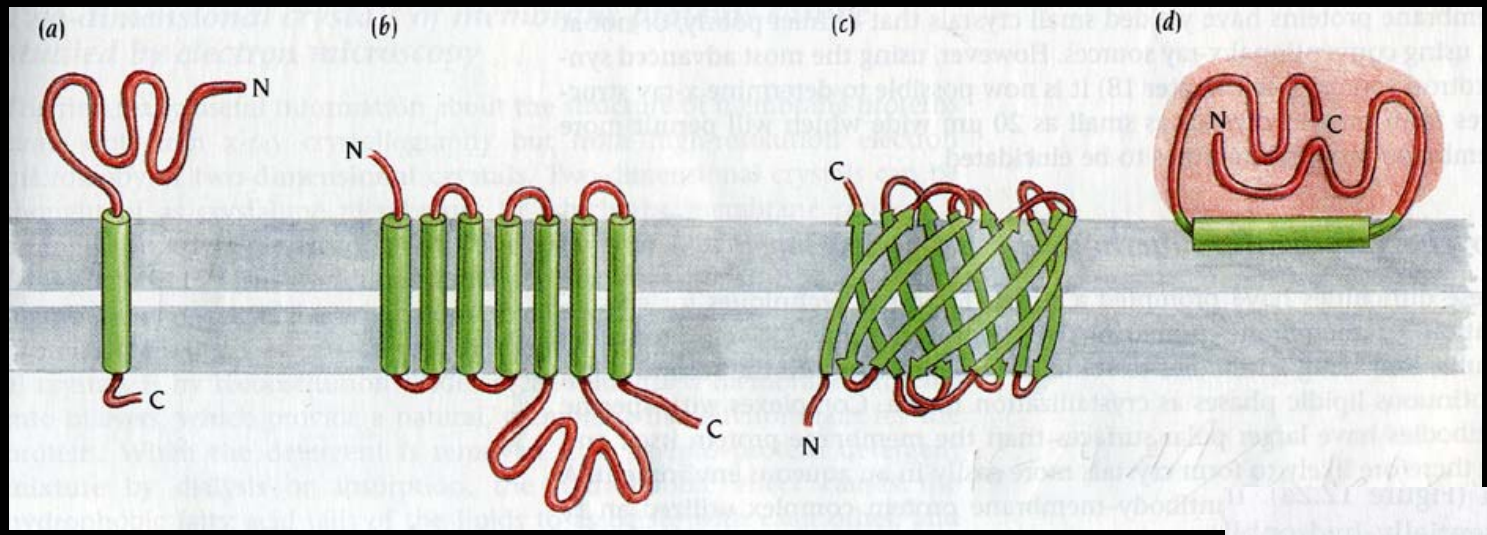


Tools that are neat

- BLAST – does the stuff you'd expect it to
  - It finds stuff.
  - There's some math about why that's good, it isn't interesting (unless you're a statistician, you aren't a statistician, right?).
  - It works, don't mess with it.
- 3DPSSM
  - What's a PSSM?
  - Whoa, 3D!
  - Does it really work?
- Trans-membrane proteins
  - 20AA α-helix and you got a transmembrane prot.
  - (see next slide)

http://www.sbg.bio.ic.ac.uk/~3dpssm/

Identify trans-membrane proteins



http://www.cbs.dtu.dk/services/SignalP/

Nobel Prize for Signal Peptides: The 1999 Nobel Prize in Physiology or Medicine has been awarded to Günter Blobel for the discovery that "*proteins have intrinsic signals that govern their transport and localization in the cell*." The first such signal to be discovered was the secretory signal peptide, which is the signal predicted by SignalP.

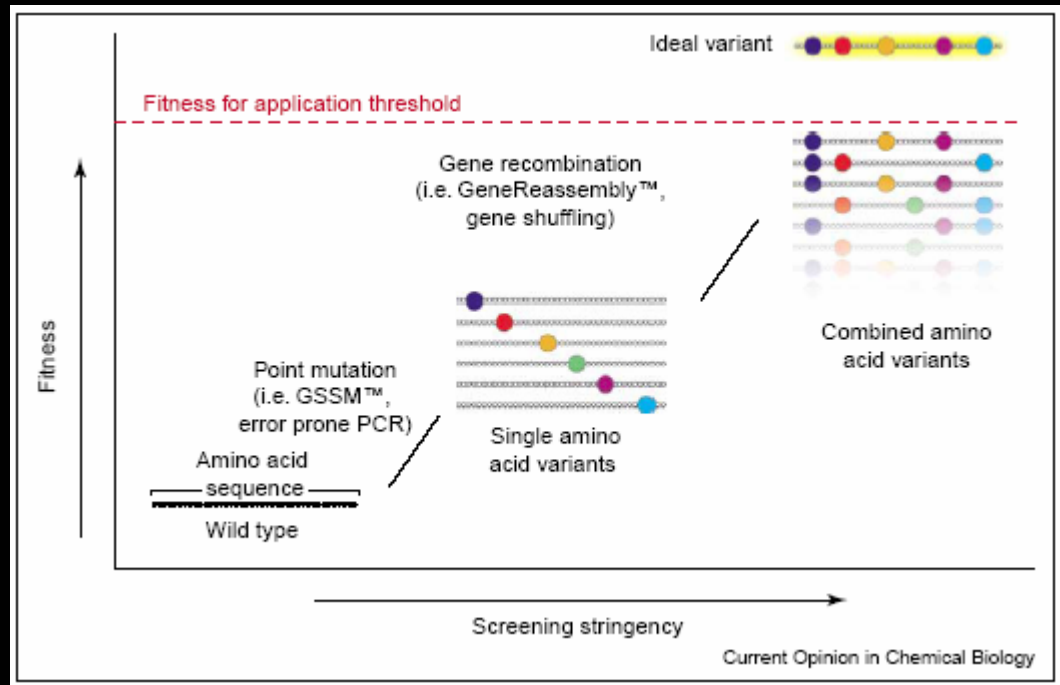
Figure 12.1 Four different ways in which protein molecules may be bound to a membrane. Membrane-bound regions are green and regions outside the membrane are red. Alpha-helices are drawn as cylinders and β strands as arrows. From left to right are (a) a protein whose polypeptide chain traverses the membrane once as an α helix, (b) a protein that forms several transmembrane α helices connected by hydrophilic loop regions, (c) a protein with several β strands that form a channel through the membrane, and (d) a protein that is anchored to the membrane by one α helix parallel to the plane of the membrane.

Three Case Studies

- Elite Organisms:
  - Single nucleotide change causes measurable phenotypic change (i.e. a fish can see different wavelengths of light), (Yokoyama *et al.* 2000, PNAS)
- Engineered Biocatalyst Proteins:
  - Diversa Corp, develops methods for high-throughput biocatalyst “discovery and optimization” (Robertson *et al.* 2004, Current Opinion in Chemical Biology)
- Two protein drugs (FDA approved):
  - TPA – Tissue Plasminogen Activator (Genentech 1986)
  - CSF – Colony Stimulating Factor (Amgen 1987)

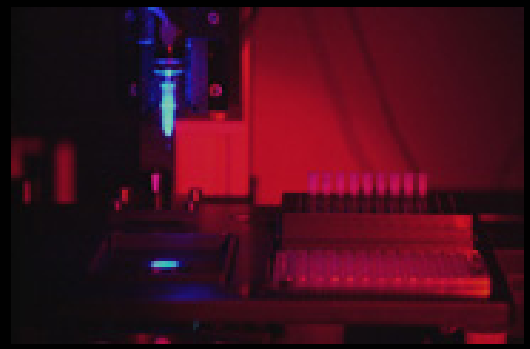
Diversa Corp and High-throughput

"Biocatalytic technologies will ultimately gain universal acceptance when enzymes are perceived to be robust, specific and inexpensive (i.e. process compatible). Genomics-based gene discovery from novel biotopes and the broad use of technologies for accelerated laboratory evolution promise to revolutionize industrial catalysis by providing highly selective, robust enzymes." (Robertson *et al.* 2004, Curr. Op. in Chem. Bio.)



Giga-Matrix Technology

GigaMatrix™ Automated Detection and Hit Recovery System



Directed Mutagenesis, Enzyme Family Classification by Support Vector Machines, and Support Vector Machines (SVMs)

given here. SVM is based on the structural risk minimization (SRM) principle from statistical learning theory.[24] In linearly separable cases, SVM constructs a hyperplane which separates two different groups of feature vectors with a maximum margin. A feature vector is represented by $\mathbf{x}_i$, with physicochemical descriptors of a protein as its components. The hyperplane is constructed by finding another vector $\mathbf{w}$ and a parameter b that minimizes $\|\mathbf{w}\|^2$ and satisfies the following conditions:

$$\mathbf{w} \cdot \mathbf{x}_i + b \geq +1, \quad \text{for } y_i = +1 \qquad \text{Group 1 (positive)} \tag{1}$$

$$\mathbf{w} \cdot \mathbf{x}_i + b \leq -1, \quad \text{for } y_i = -1 \qquad \text{Group 2 (negative)} \tag{2}$$

where y_i is the group index, $\mathbf{w}$ is a vector normal to the hyperplane, $|b|/\|\mathbf{w}\|$ is the perpendicular distance from the hyperplane to the origin and $\|\mathbf{w}\|^2$ is the Euclidean norm of $\mathbf{w}$. After the determination of $\mathbf{w}$ and b, a given vector $\mathbf{x}_i$ can be classified by:

$$sign[(\mathbf{w} \cdot \mathbf{x}) + b] \tag{3}$$

(Cai, Proteins, 2004)

In nonlinearly separable cases, SVM maps the input variable into a high dimensional feature space using a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$. An example of a kernel function is the Gaussian kernel which has been extensively used in different studies:[17,24–27,29–31]

$$K(\mathbf{x}_i,\mathbf{x}_j) = e^{-\|\mathbf{x}_j - \mathbf{x}_i\|^2/2\sigma^2} \tag{4}$$

Based on earlier study[27,38] and our own analysis, Gaussian kernel function seems to produce better results than other kernel functions. Linear support vector machine is applied to this feature space and then the decision function is given by:

$$f(\mathbf{x}) = sign\left(\sum_{i-1}^{l} \alpha_i^0 y_i K(\mathbf{x}, \mathbf{x}_i) + b\right) \tag{5}$$

where the coefficients α_i^0 and b are determined by maximizing the following Langrangian expression:

$$\sum_{i-1}^{l} \alpha_i - \frac{1}{2} \sum_{i-1}^{l} \sum_{j-1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{6}$$

under conditions:

$$\alpha_i \geq 0 \quad \text{and} \quad \sum_{i-1}^{l} \alpha_i y_i = 0 \tag{7}$$

A positive or negative value from Eq. (3) or Eq. (5) indicates that the vector $\mathbf{x}$ belongs to the positive or negative group respectively.

Vapnick, V. (1995)
The Nature of Satistical Learning Theory. Springer, New York.

Legal Problems with BioTech:
Why this is a huge enterprise

- Approaches to drug patenting:
  - Composition of Matter
  - Process Patent (i.e. especially with FDA approval)
  - Structure Characterization
  - Use Patent
- FDA Approval
  - Takes years and years
  - A main reason why it takes so long for a BioTech firms to return on investment (i.e. target buyouts before product)

Goals

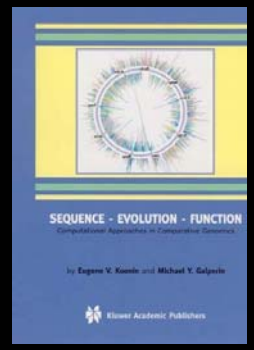
- Introduce some current issues
- Introduce several resources that are around to address those problems
- “I was a teenage genetic engineer”; “Temper the arrogance”
  - On DNA Polymerase:

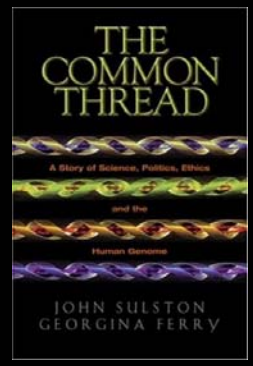
    “Because the complexity of polymerization reactions in vitro pales in comparison to the enormous complexity of multiple, highly integrated DNA transactions in cells, the biggest challenge of all may be to use our biochemical understanding of replication fidelity to reveal, and perhaps even predict, biological effects. In this regard, *any arrogance about our current level of understanding should be tempered* by the realization that the number of template-dependent DNA polymerases encoded by the human genome may be more than twice that suspected only four years ago.” (Kunkel and Bebenek, Annu. Rev. Biochem., 2000)

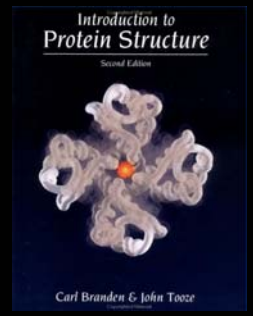
Reading

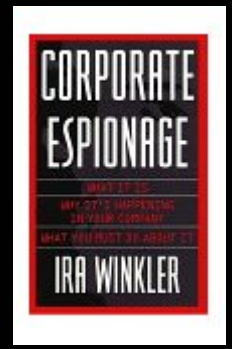
- Eugene Koonin:
  - Sequence - Evolution - Function: Computational Approaches in Comparative Genomics (2002)
- John Sulston:
  - The Common Thread: A Story of Science, Politics, Ethics and the Human Genome (2002)
- Branden & Tooze:
  - Introduction to Protein Structure (1999)
- Ira Winkler:
  - Corporate Espionage (1997)
  - Spies Among Us: The Spies, Hackers, and Criminals Who Cost Corporations Billions (2004)
- Presentations from the O'Reilly BioCon 2003:

  $ wget -r -A ppt,pdf http://conferences.oreillynet.com/cs/bio2003/view/e_sess/3516









Acknowledgements

- Co-workers: John B, Kristin W, Eric D, and others.
- O'Reilly Bioinformatics Con 2003
- Some other people.