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

21st Chaos Communication Congress
December 27th to 29th, 2004
Berliner Congress Center, Berlin, Germany

Russell Hanson
Dec 27, 2004
Outline

- Analogies – why this talk?
- 2600 article – transgenes
- Engineering proteins
- Computer tools for genome analysis
- Conclusions
The Analogy

Instruction Pointer : Machine Code :: Ribosome : RNA

Stack Pointer (SP) の動き

5 Å Map Of The Large Ribosomal Subunit
The Analogies, cont.

Instruction Pointer : Machine Code ::
Ribosome : RNA

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


Figure 1: The subunits of a ribosome. Side and front view.
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, …
Hacking the Genome

by Professor L

2600 Magazine | Winter 2003-2004

The creation of plasmids (or other vectors) with transgenic payloads is made possible by the existence of DNA splicing enzymes. Simple laboratory techniques allow the extraction of naturally occurring plasmids from bacteria and splicing the DNA for the new gene into them. The hardest part is figuring out which combination of genes to insert into a host in order to get a desired effect. However, these techniques are beyond the scope of this introductory article. For our purposes, we can just buy plasmids with our desired genes from a scientific supply house. An E. coli plasmid with the Luciferase gene in it is called pUC18-luxR, and can be purchased from many places (see sources section, below).

Folding of Firefly Luciferase

Firefly Photinus pyralis is a popular model protein used in studies of intramolecular and extramolecular folding. To provide a physical context, we have started an in-vitro study of luciferase folding. We have identified several unfolding intermediates and investigated their role in the reassembly of the enzyme.

Crystal structure: Conti et al., (1996) Structure 4, 237; PDB entry 1LCI

References:


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.

Noble Prize 1984 Bruce Merrifield: solid phase peptide synthesis
PCR lets you assemble pieces ad infinitum

- Sketch:

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

- DNA
- mRNA
- transcription factors
- splicing factors
- Receptors
- Enzymes
  - structural proteins
  - structural sugars
  - structural lipids
- signaling molecules
- environment
- other cells
- creation
- regulation
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$_\alpha$) 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$\equiv$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$_\alpha$ 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$_\alpha$ 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) design/architecture of systems
- 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 $\varphi$ and $\psi$, in allowable regions
- given a protein of 150 amino acids
- $= 3^{150}$ possible structures $\approx 10^{68}$
- time of bond rotation $10^{-12}$s
- $10^{68} \times 10^{-12}$s = $10^{56}$ sec = $10^{48}$ years
- Real folding times are 0.1 – 1000 sec

Life on earth $3.8 \times 10^9$ years
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 (α4) and β-sheet (β4) bundles, ββα- (βα) or more complex arrangements, e.g., (β3α2), are molecules with a built-in pathway for folding.
“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.”


**Def: Ligand**

Any molecule that binds specifically to a receptor site of another molecule; proteins embedded in the membrane exposed to extracellular fluid.
One way to test for ligand binding

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

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.”
Bioinformatics Databases

- Completely sequenced genomes
- COG – Clusters of orthologous groups
- NR@ncbi
- Pfam
- SwissProt
- SMART

BLAST with CD ũ-on (Conserved Domain)

PSI-Blast searches the Non-redundant (NR) database

21C3 – Berlin
How to Access the Human Genome (and other sequenced genomes)


<table>
<thead>
<tr>
<th>Directory</th>
<th>Date</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>A thaliana</td>
<td>10/17/2003</td>
<td>0:00:00</td>
</tr>
<tr>
<td>Anopheles gambiæ</td>
<td>5/7/2002</td>
<td>0:00:00</td>
</tr>
<tr>
<td>Bacteria</td>
<td>4/7/2004</td>
<td>18:33:00</td>
</tr>
<tr>
<td>C elegans</td>
<td>6/14/2002</td>
<td>0:00:00</td>
</tr>
<tr>
<td>D melanogaster</td>
<td>10/19/2000</td>
<td>0:00:00</td>
</tr>
<tr>
<td>H sapiens</td>
<td>4/15/2004</td>
<td>0:23:00</td>
</tr>
<tr>
<td>Leptospira interrogans serovar Copenhageni</td>
<td>3/22/2004</td>
<td>17:49:00</td>
</tr>
<tr>
<td>M musculus</td>
<td>11/2/1999</td>
<td>0:00:00</td>
</tr>
<tr>
<td>P falciparum</td>
<td>5/12/2002</td>
<td>0:00:00</td>
</tr>
<tr>
<td>Plasmodium falciparum</td>
<td>5/11/1999</td>
<td>0:00:00</td>
</tr>
<tr>
<td>Flies</td>
<td>10/11/2002</td>
<td>0:00:00</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>File</th>
<th>Size</th>
<th>Date</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>hs_phase0.fna.gz</td>
<td>94408 KB</td>
<td>4/14/2004</td>
<td>21:56:00</td>
</tr>
<tr>
<td>hs_phase1.fna.gz</td>
<td>606370 KB</td>
<td>4/14/2004</td>
<td>22:35:00</td>
</tr>
<tr>
<td>hs_phase2.fna.gz</td>
<td>48487 KB</td>
<td>4/14/2004</td>
<td>22:39:00</td>
</tr>
<tr>
<td>hs_phase3.fna.gz</td>
<td>1090520 KB</td>
<td>4/15/2004</td>
<td>0:23:00</td>
</tr>
</tbody>
</table>


- hs_phs0.fna.gz Survey sequence (appx 200x coverage)
- hs_phs1.fna.gz Unordered contigs (each contig is appx 200x coverage)
- hs_phs2.fna.gz Ordered contigs (each contig is appx 200x coverage)
- hs_phs3.fna.gz Finished sequence
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](http://www.accelrys.com/insight/consensus.html)) to find conserved regions

- **5th Step:** Predict secondary structure: [http://www.compbio.dundee.ac.uk/~www-jpred/](http://www.compbio.dundee.ac.uk/~www-jpred/)
  - **Prediction method:** “J net; 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


• 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

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

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

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

- Rooted vs Unrooted Trees
  - Cladograms and phylogenies 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.
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)
“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. 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:

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

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

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:

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

(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:

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

Based on earlier study 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:

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

where the coefficients \( \alpha_i^0 \) 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.
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 resources that address some of those issues
• “I was a teenage genetic engineer”
  – 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:
- Branden & Tooze:
  - Introduction to Protein Structure (1999)
- Ira Winkler:
  - Corporate Espionage (1997)
- Presentations from the O’Reilly BioCon 2003:
  
  $ wget -r -A ppt.pdf
Acknowledgements

• GIT co-workers: John B, Kristin W, Eric D
• O’Reilly Bioinformatics Con 2003
• Some other people.