Bioinformatics (and Constraints)

David Gilbert

City University

www.soi.city.ac.uk/~drg

drg@soi.city.ac.uk

Based on material jointly developed with

Rolf Backofen

Ludwig-Maximillians University, Munich, Germany

backofen@informatik.uni-muenchen.de

www.tcs.informatik.uni-muenchen.de/~backofen

Contents

- Biological background
- The Central Dogma
- Overview of Bioinformatics
- Current problem areas
- Resources

What is Bioinformatics

- *Bio* - molecular biology
- *Informatics* - computer science.
- BioInformatics - solving problems arising from biology using methodology from computer science.
  (Computational Biology - USA).

Related but different...

Apply principles from biology to derive novel approaches in computer science:

- biocomputing
- neural computing
- genetic algorithms
- evolutionary computing

Human genome sequenced! 23 June 2000

- "The most wondrous map ever produced by human kind"
- Scientists jointly announced that they had obtained a near complete set of the biochemical instructions for human life.
- "One of the most significant scientific landmarks of all time, comparable with the invention of the wheel or the splitting of the atom"

An advance or ???

- The genetic information will revolutionise medicine over the coming decades, giving us new tests and drugs for previously untreatable diseases.
- Publicly and privately funded researchers
- Human Genome Project, immediately makes all of its data freely available on the net
- Dr Craig Venter, head of Celera Genomics, intends to patent some of its discoveries & to sell its information to drug companies.

Issues

- How will this benefit humanity
- Genetically modified crops - contamination escapes...
- Genetically modified food - ok?
- Genetically modified wine....!
- Genes & behaviour - really?
- testing on animals - why?
- Gene therapy - benefits outweigh dangers?

What's in the draft genome sequence?

- **Features found in DNA & their structure**
  - Genes (coding regions & their control regions)
  - Junk DNA (simple repeats, "dead" viruses, pseudo-genes)
- **How to hunt for genes**
  - Homology methods - compare DNA sequence to database of known genes (high accuracy)
  - *Ab initio* prediction (low accuracy)
  - Experimental methods (high accuracy)
- **Bioinformatics - putting it all together**
  - Annotation - combining expert knowledge & algorithms
  - Visualization - picture paints 1000 words
  - interconnecting databases

Bioinformatics is about:

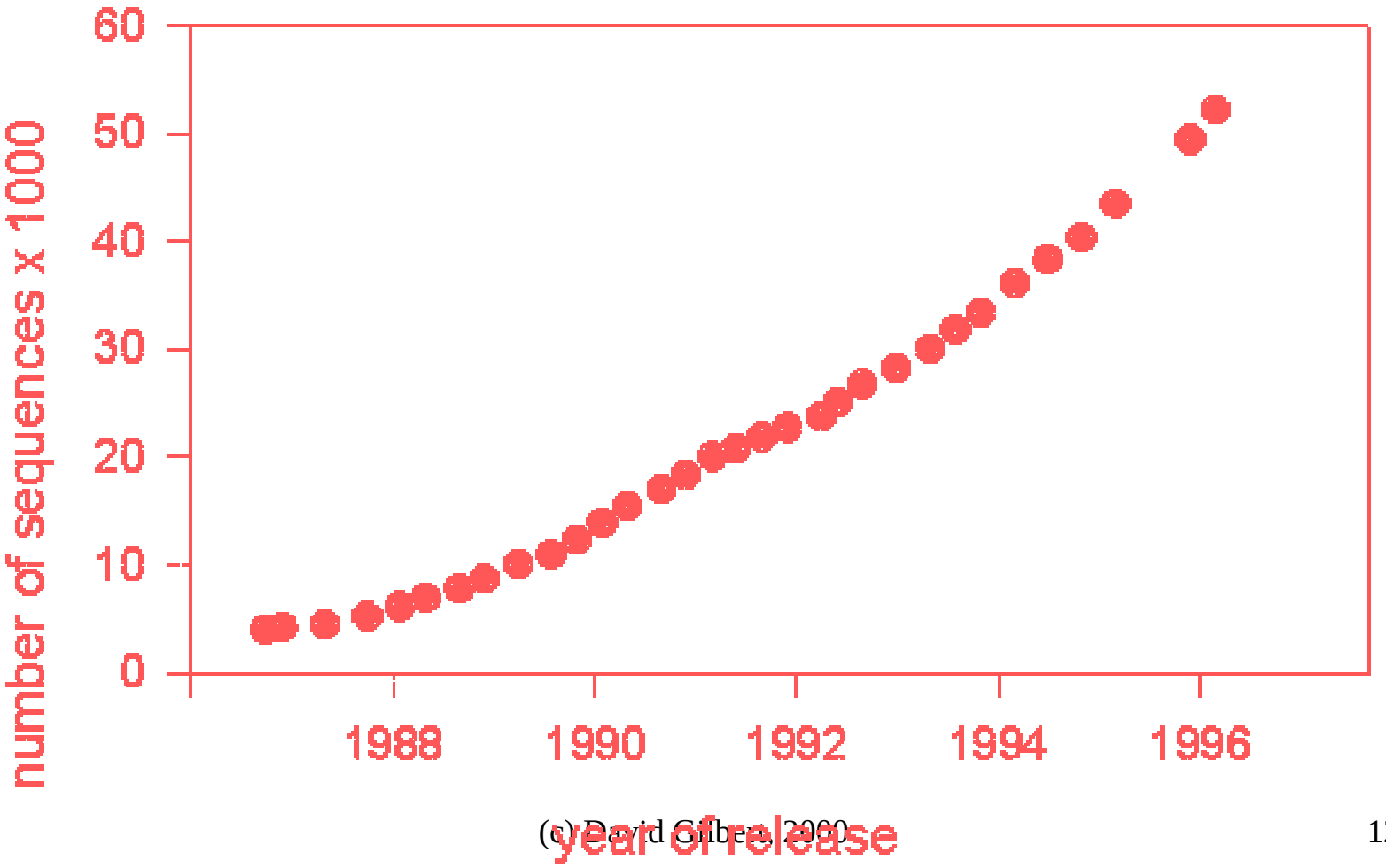
- Elicitation of DNA sequences from genetic material
- Sequence annotation (e.g. with information from experiments)
- Understanding the control of gene expression (i.e. under what circumstances proteins are transcribed from DNA)
- The relationship between the amino acid sequence of proteins and their structure.

Aim of research in Bioinformatics

Understand the functioning of living things - to “improve the quality of life”.

- drug design
- identification of genetic risk factors
- gene therapy
- genetic modification of food crops and animals, etc.
- (biological warfare, crime etc).

Flood of data! (SWISSPROT)



How can we analyse the flood of data ?

- Data: don't just store it, analyze it ! By comparing sequences, one can find out about things like

  - ancestors of organisms
  - phylogenetic trees
  - protein structures
  - protein function

Using the genome

genetic information

↓ molecular dynamics

molecular structure

↓ biophysics

biochemical function

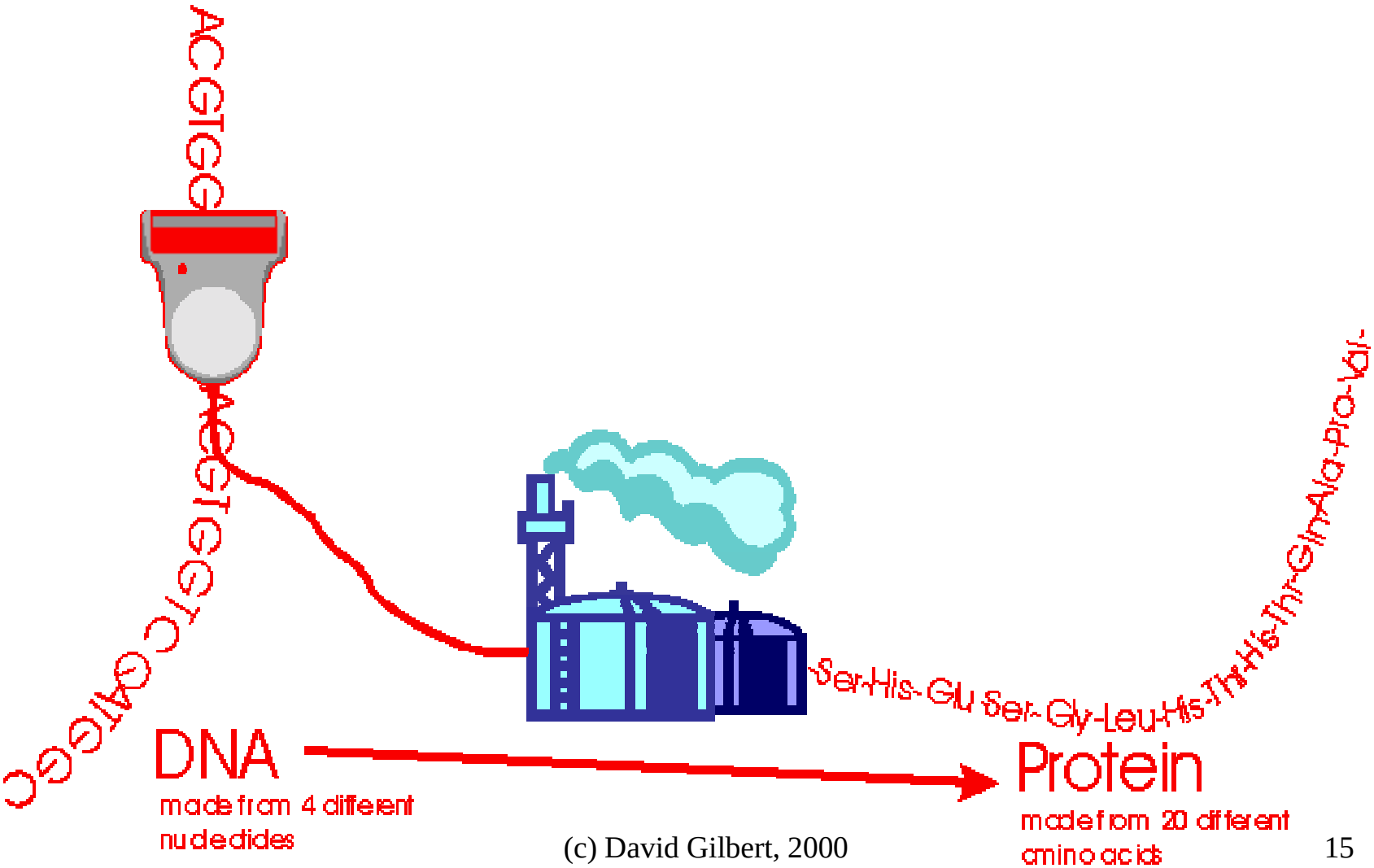
↓ biochemistry

biological behaviour

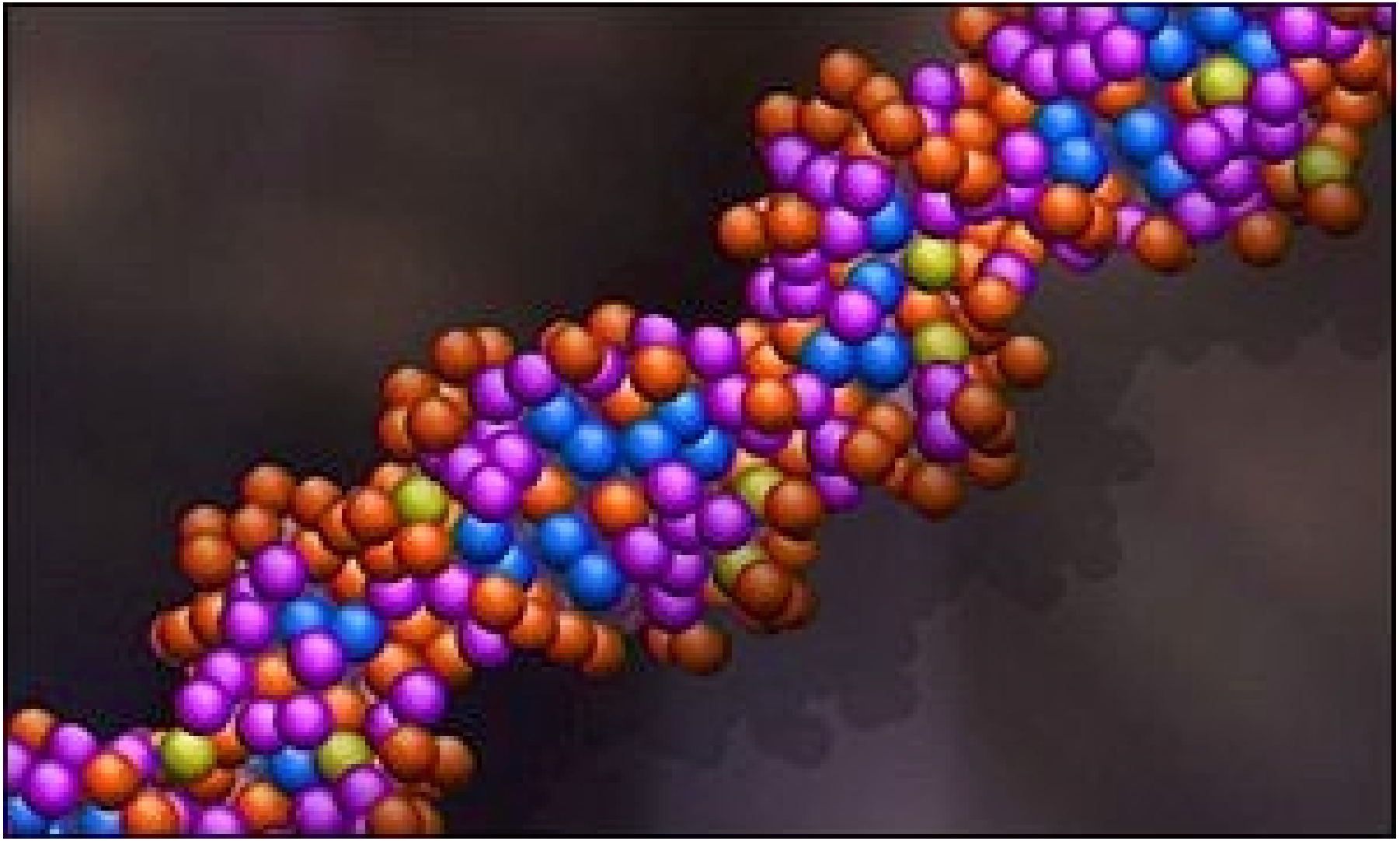
- redundancy in genetic information
- single genes have multiple functions
- genes 1-D, gene products 3-D

Molecular biology: flow of information

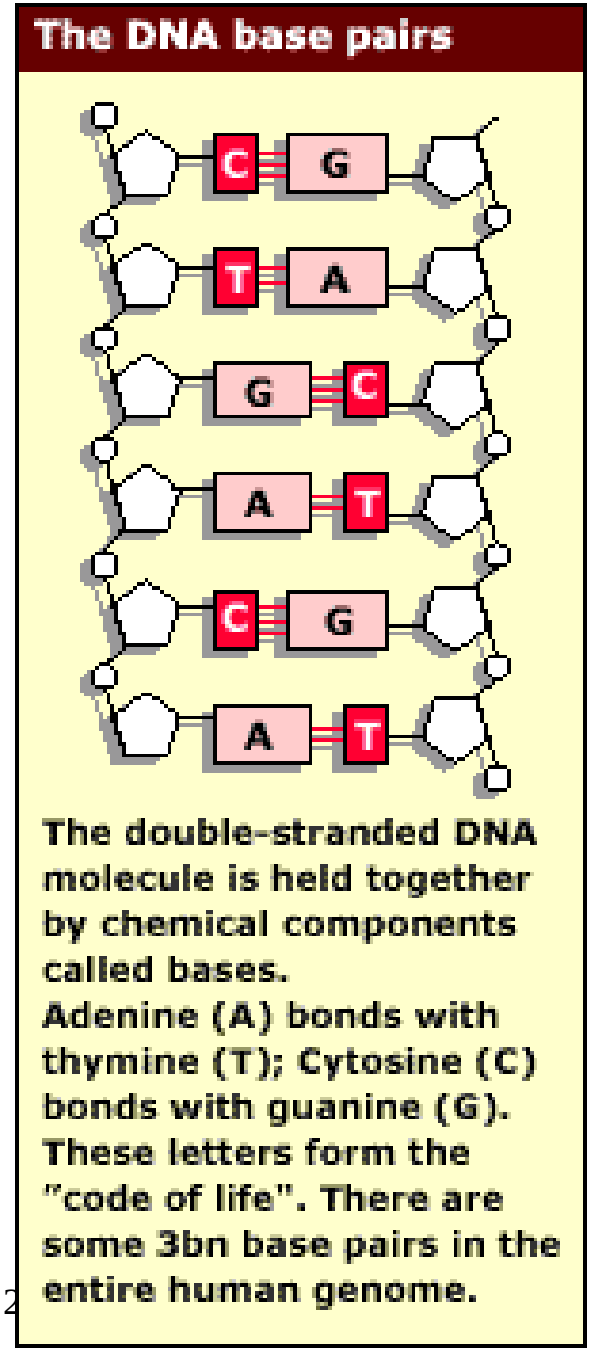
DNA → RNA → Protein → Function



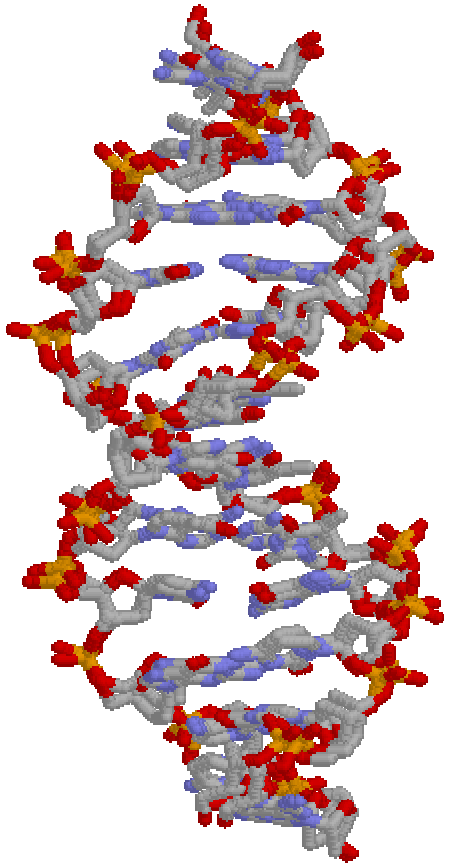
DNA double helix



DNA base-pairs



DNA

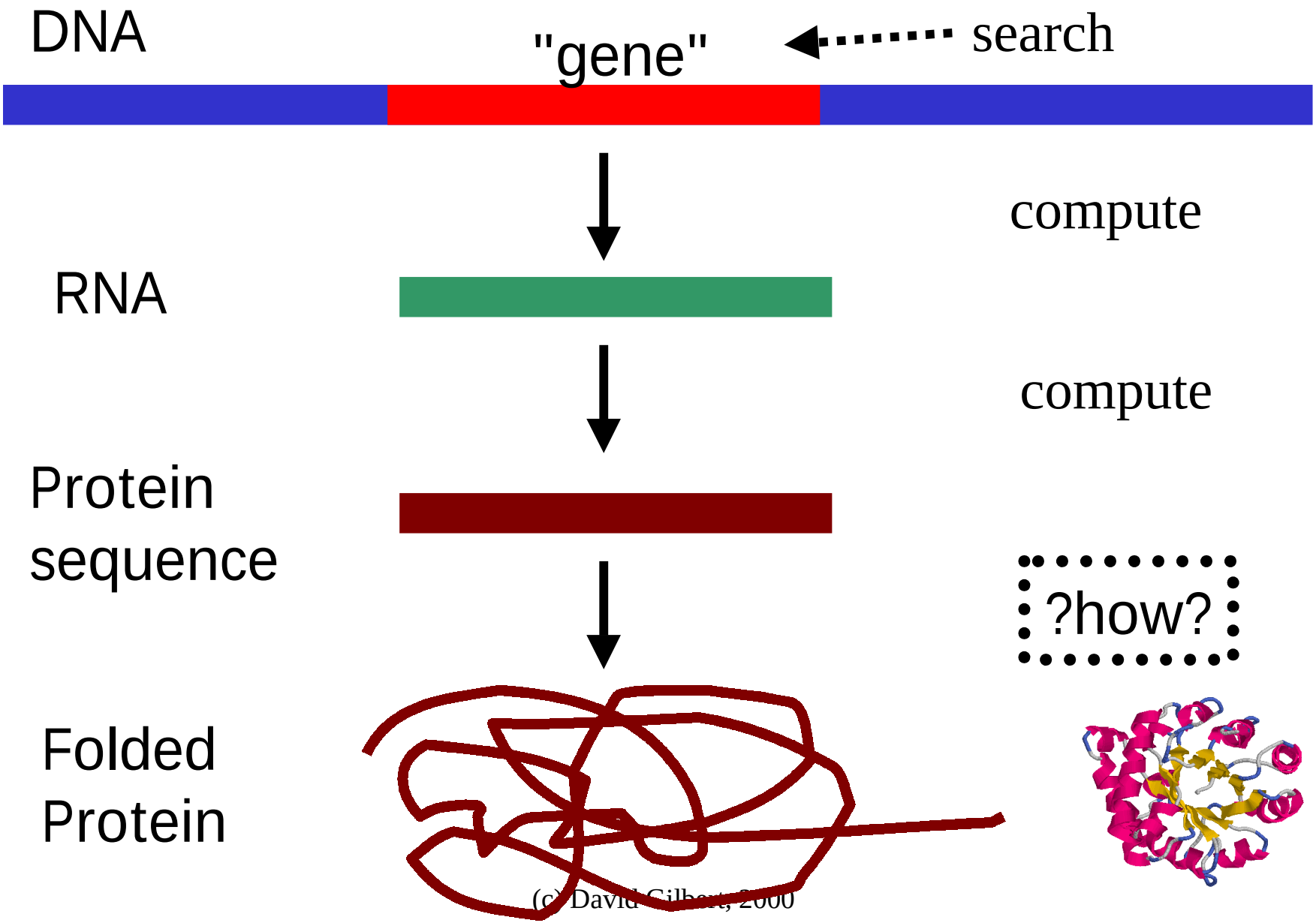


AAAAGAAAAGGTTAGAAAGATGAGAGATGATAAAGGGTCCATTTG
AGGTTAGGTAATATGGTTTGGTATCCCTGTAGTTAAAAGTTTTTG
TCTTATTTTAGAATACTGTGACTATTTCTTTAGTATTAATTTTTC
CTTCTGTTTTCCTCATCTAGGGAACCCCAAGAGCATCCAATAGAA
GCTGTGCAATTATGTAAAATTTTCAACTGTCTTCCTCAAAATAAA
GAAGTATGGTAATCTTTACCTGTATACAGTGCAGAGCCTTCTCAG
AAGCACAGAATATTTTTATATTTCCTTTATGTGAATTTTTAAGCT
GCAAATCTGATGGCCTTAATTTCCTTTTTGACACTGAAAGTTTTG
TAAAAGAAATCATGTCCATACACTTTGTTGCAAGATGTGAATTAT
TGACACTGAACTTAATAACTGTGTACTGTTCGGAAGGGGTTCCTC
AAATTTTTTGACTTTTTTTGTATGTGTGTTTTTTCTTTTTTTTTA
AGTTCTTATGAGGAGGGAGGGTAAATAAACCACTGTGCGTCTTGG
TGTAATTTGAAGATTGCCCCATCTAGACTAGCAATCTCTTCATTA
TTCTCTGCTATATATAAAACGGTGCTGTGAGGGAGGGGAAAAGCA
TTTTTCAATATATTGAACTTTTGTACTGAATTTTTTTGTAATAAG
CAATCAAGGTTATAATTTTTTTTAAAATAGAAATTTTGTAAGAAG
GCAATATTAACCTAATCACCATGTAAGCACTCTGGATGATGGATT
CCACAAAACTTGGTTTTATGGTTACTTCTTCTCTTAGATTCTTAA
TTCATGAGGAGGGTGGGGGAGGGAGGTGGAGGGAGGGAAGGGTTT
CTCTATTAAAATGCATTCGTTGTGTTTTTTAAGATAGTGTAACTT
GCTAAATTTCTTATGTGACATTAACAAATAAAAAAGCTCTTTTAA
TATTAGATAA

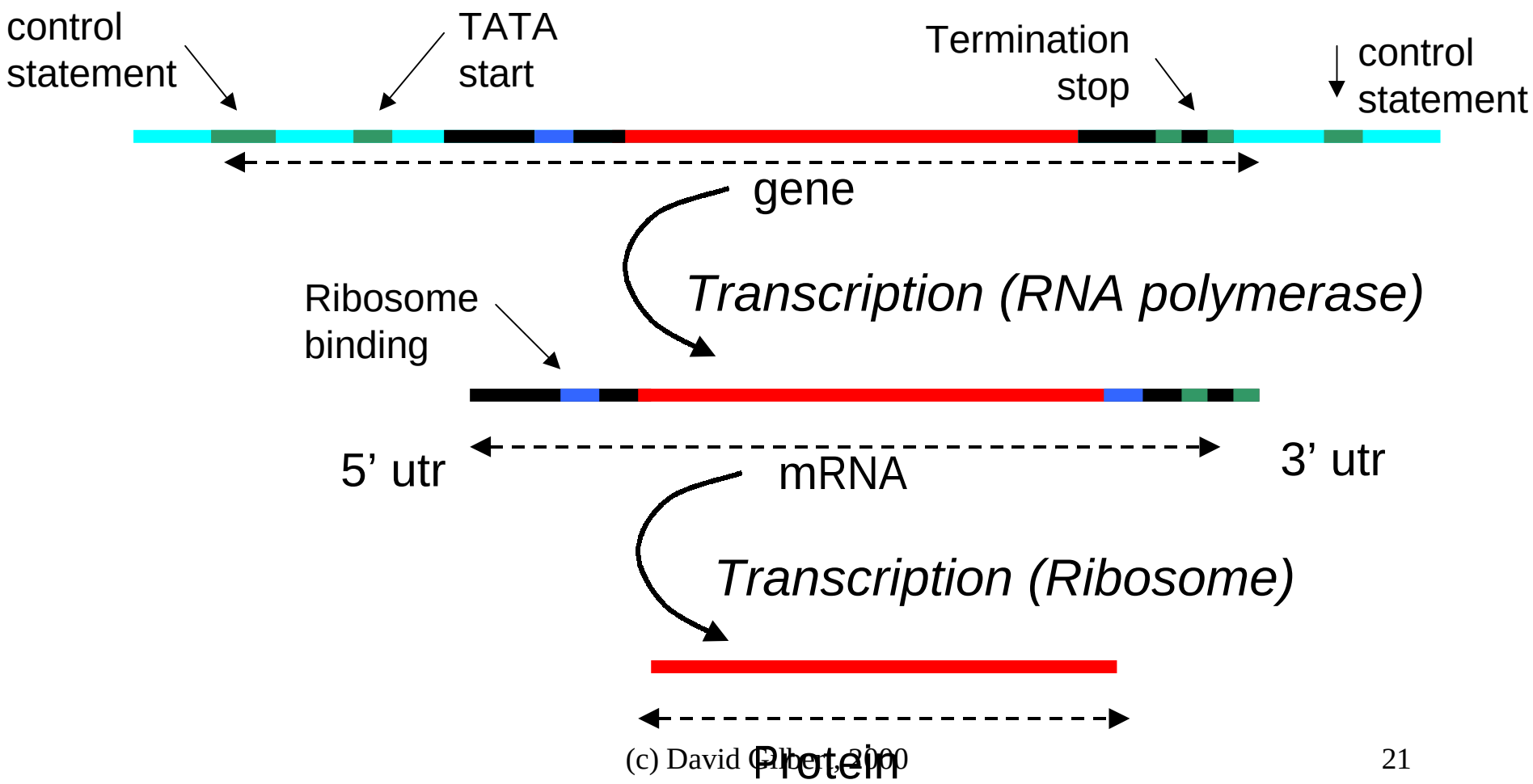
Some facts...

- DNA differs between humans by 0.2%, (1 in 500 bases).
- Human DNA is 98% identical to that of chimpanzees.
- 97% of DNA in the human genome has no known function.
- 3.10^9 letters in the DNA code in every cell in your body.
- 10^{14} cells in the body.
- 12,000 letters of DNA decoded by the Human Genome Project every second.

Gene structure



DNA (gene) → RNA → Protein



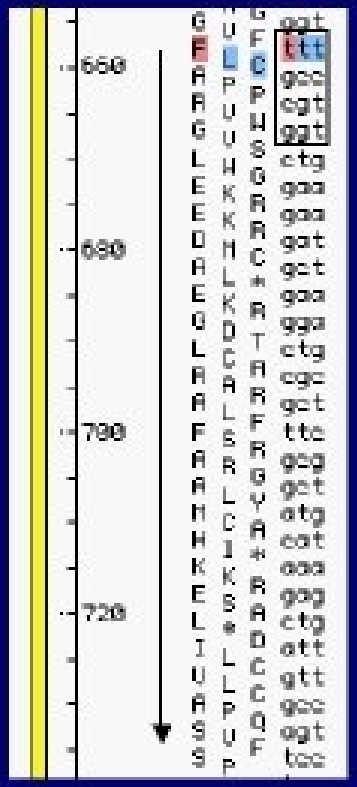
Numbers of genes

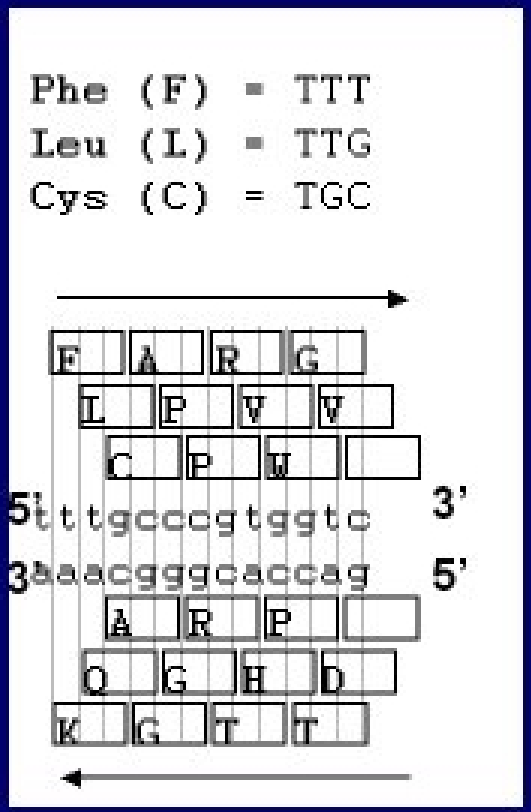
- humans & mice: 60,000 - 100,000
- C. elegans (worm): 19,000
- S. cerevisiae (yeast): 6,000
- Tuberculosis microbe: 4,000

Genetic Code: 3 bases = 1 amino acid

First position (5' end)	*Second position*				Third position (3' end)
	T	C	A	G	
T	Phe	Ser	Tyr	Cys	T
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	T
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	T
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	T
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

DNA is translated into protein in one of 6 reading frames





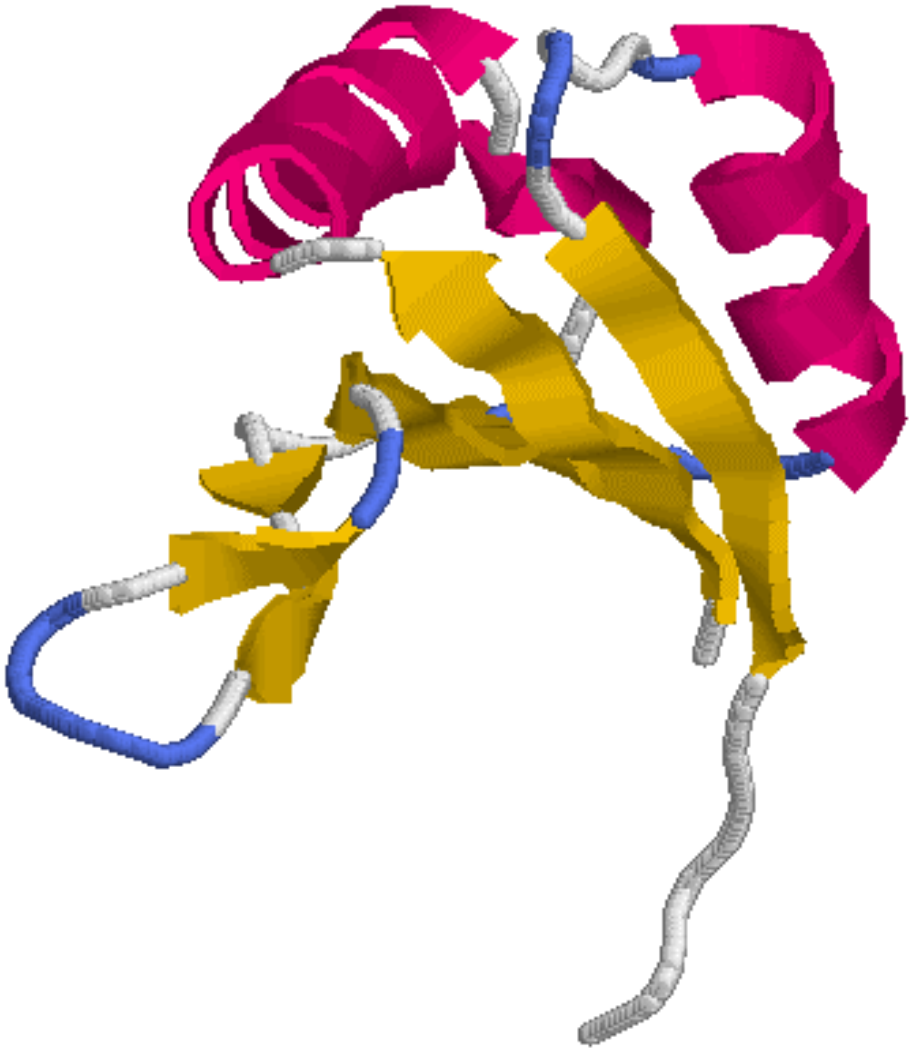
Nucleotide sequence

```
SQ   Sequence 1344 BP; 291 A; 374 C; 401 G; 278 T; 0 other;
gcacgagtaa acatgcactt cccaggccac agcagcaaga aggaggaatc tgcccaagcg
gccctcacga agctgaactc ctggttcccc accaccaaga accccgtcat catcagcgcc
cccatgtatc tcatcgccaa cggcactctt gcggccgagg tatccaaggc cggcggtatt
ggctttgtcg ccggcggctc cgacttccgc cccggctcct cccacctaac cgccctctct
accgaactcg cctccgcccg cagccgcctc ggtcttaccg accgccccct cacccctctc
cccggcattg gcgtcggcct cattttaacc cacaccatct ccgttcccta cgtaaccgac
accgtcctgc ccatcctgat cgaacactcc ccgcaagcag tctggctctt cgccaacgac
ccggatttcg aggcctcttc cgagcctggc gcaaagggaa cagcaaagca aatcatcgag
gcccttcacg cttcggggtt cgtggtattc tttcaagtag gcacggtgaa agatgcaagg
aaggcggcgg cagatggggc agatgtgatt gttgcgcaag ggatcgatgc gggagggcat
cagcttgcta cagggagtgg gattgtgagt ttggtaccgg aggttaggga tatgcttgat
agagagttca aggaacgaga ggtggtggtt gtggcggcgg gaggtgtggc ggatgggagg
ggggttgtag gggcgctggg tctaggcgcc gagggtgtgg tattgggtac taggttcacc
gtagcagtcg aagcttccac ccccgagttc cgcaggaagg tcatcctcga gacaaacgat
ggtggtctca acaccgtcaa atcccatttc cacgaccaaa tcaactgcaa cacaatctgg
cacaacgtct acgacgggcg agccgttcgc aatgcctcct acgacgacca cgcggccggt
gtccccttg aagagaatca caagaagttc aaggaggcag cgagctctgg ggataactcg
cgggctgtga cttggtccgg gactgctgtg ggtctgataa aggaccagag gccggctggc
gatattgtta gggagttgag ggaagaggcc aaagagagga tcaagaagat tcaggctttt
gctgcttaag ggggggccta aggggtgccg cgtgtaatga tgggtgattg aaaacgcatg
ggtcaatatc gtaactacag atcgcaagcg agtttggtct tcggttcctt ggtgatcttt
gactgtgttc tgcctcttta ttgctcttcg tcgtaatggg cacgagggat gggaagcaaa
aacatgataa ttcgaactcg tgcc
```

Protein (amino acid) sequence

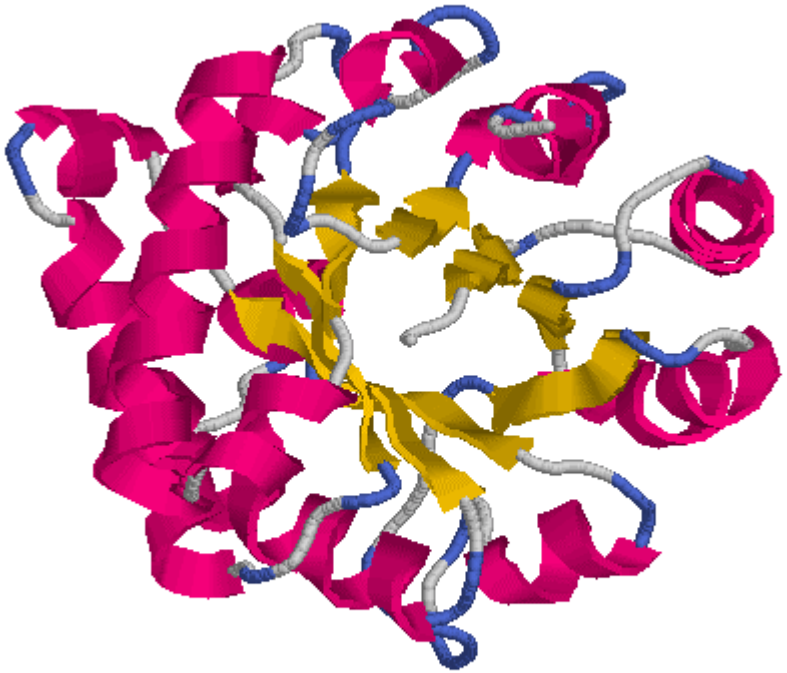
```
DR   EMBL; U22530; AAA64218.1; -.
DR   HSSP; P03122; 2BOP.
KW   Oxidoreductase; Dioxygenase; Flavoprotein; FMN.
FT   PROPEP        1     15       POTENTIAL.
FT   CHAIN        16    378       2-NITROPROPANE DIOXYGENASE.
SQ   SEQUENCE   378 AA;  39916 MW;  E453EB43FD23E441 CRC64;
MHFPGHSSKK EESAQAALTK LNSWFPTTKN PVIISAPMYL IANGTLAAEV SKAGGIGFVA
GGSDFRPGSS HLTALSTELA SARSRLGLTD RPLTPLPGIG VGLILTHTIS VPYVTDTVLP
ILIEHSPQAV WLFANDPDFE ASSEPGAKGT AKQIIEALHA SGFVVFFQVG TVKDARKAAA
DGADVIVAQG IDAGGHQLAT GSGIVSLVPE VRDMLDREFK EREVVVVAAG GVADGRGVVG
ALGLGAEGVV LGTRFTVAVE ASTPEFRRKV ILETNDGGLN TVKSHFHDQI NCNTIWHNVY
DGRAVRNASY DDHAAGVPFE ENHKKFKEAA SSGDNSRAVT WSGTAVGLIK DQRPAGDIVR
ELREEAKERI KKIQAFAA
```

Protein structure



Plait

Protein structure

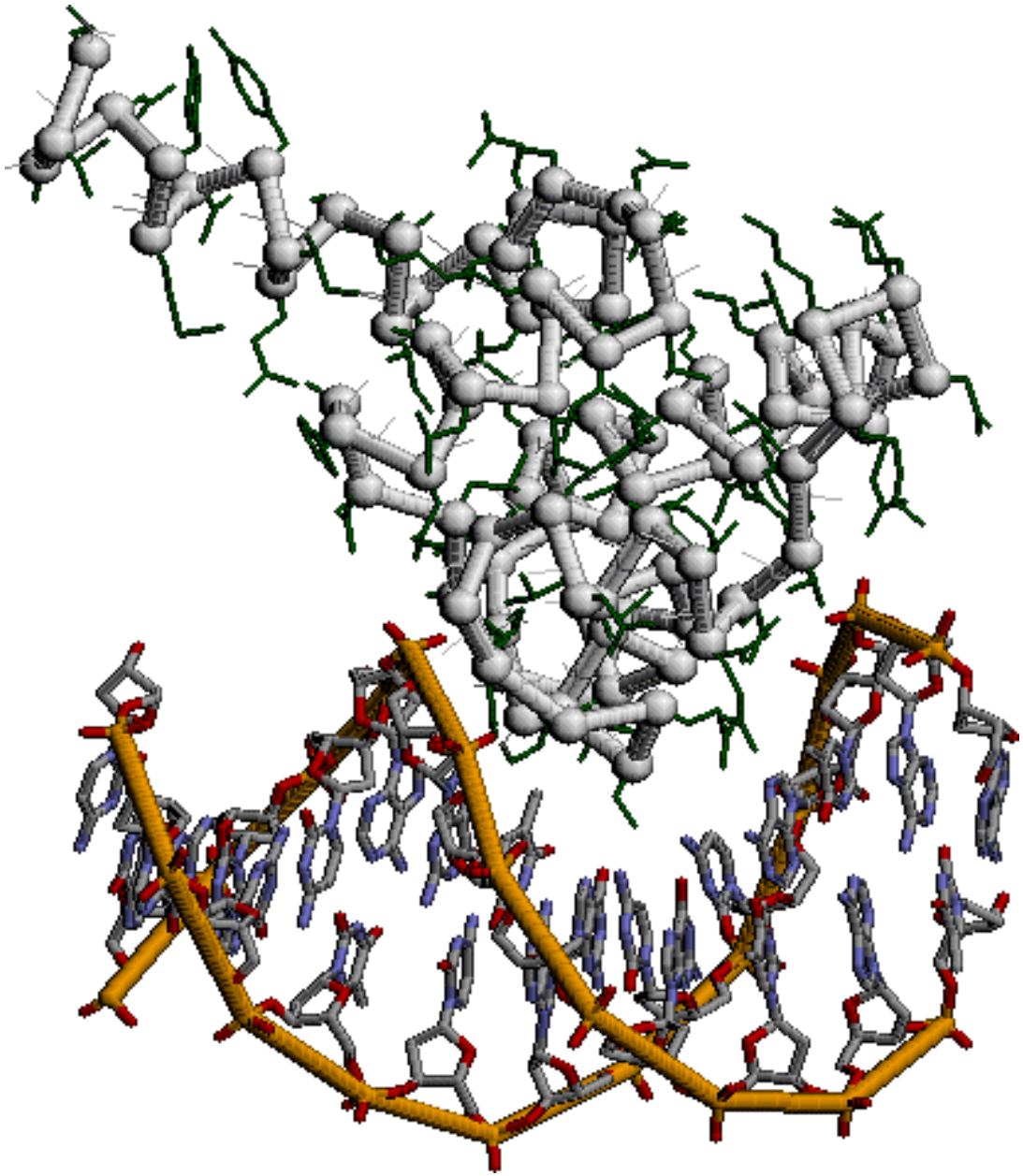


TIM barrel

Proteins

- ~ 60% of dry mass of a living cell
- Linear heteropolymers
- Constructed from chain of amino acids (20 different types)
- Function of proteins (and RNA) determined by their structure,
- structure uniquely determined by the sequence of amino acids, (RNA: nucleotides).

DNA-protein interaction



Human genetic variations (Single Nucleotide Polymorphisms)

- SNP's - "genetic indivuality"
- ~ 1/1000 bases variable (2 humans)
- Make us more/less susceptible to diseases
- May influence the effect of drug treatments

```
TTTGCTCCGTTTTCA
TTTGCTCYGTTTTCA
TTTGCTCTGTTTTCA
```

HUMAN INDIVIDUALITY

The Sanger Centre

SNP implicated in coronary disease

LDL gene sequence

```
TTT TAC GGC ATC
Phe Tyr Asn Met
```

```
TTT TAC GTC ATC
Phe Tyr Ser Met
```

Associated with
high cholersterol

The Central Dogma of information flow in biology

The sequence of amino acids making up a protein and hence its structure (folded state) and thus its function, is determined by transcription from DNA via RNA.

A Holy Grail

- Develop computational methods to determine protein structure from amino-acid sequence.

3 main classes of problem areas

- Central Dogma related: sequence, structure or function
- Data related: storage, retrieval & analysis (exponential growth of knowledge in molecular biology)
- Simulation of biological processes - protein folding (molecular dynamics) or metabolic pathways

Current problem areas

- Physical map
- Alignment & threading
- Protein structure prediction
- Search and pattern discovery
- Phylogenetic trees
- Metabolic pathways & regulatory networks

Protein folding problem (Structure prediction)

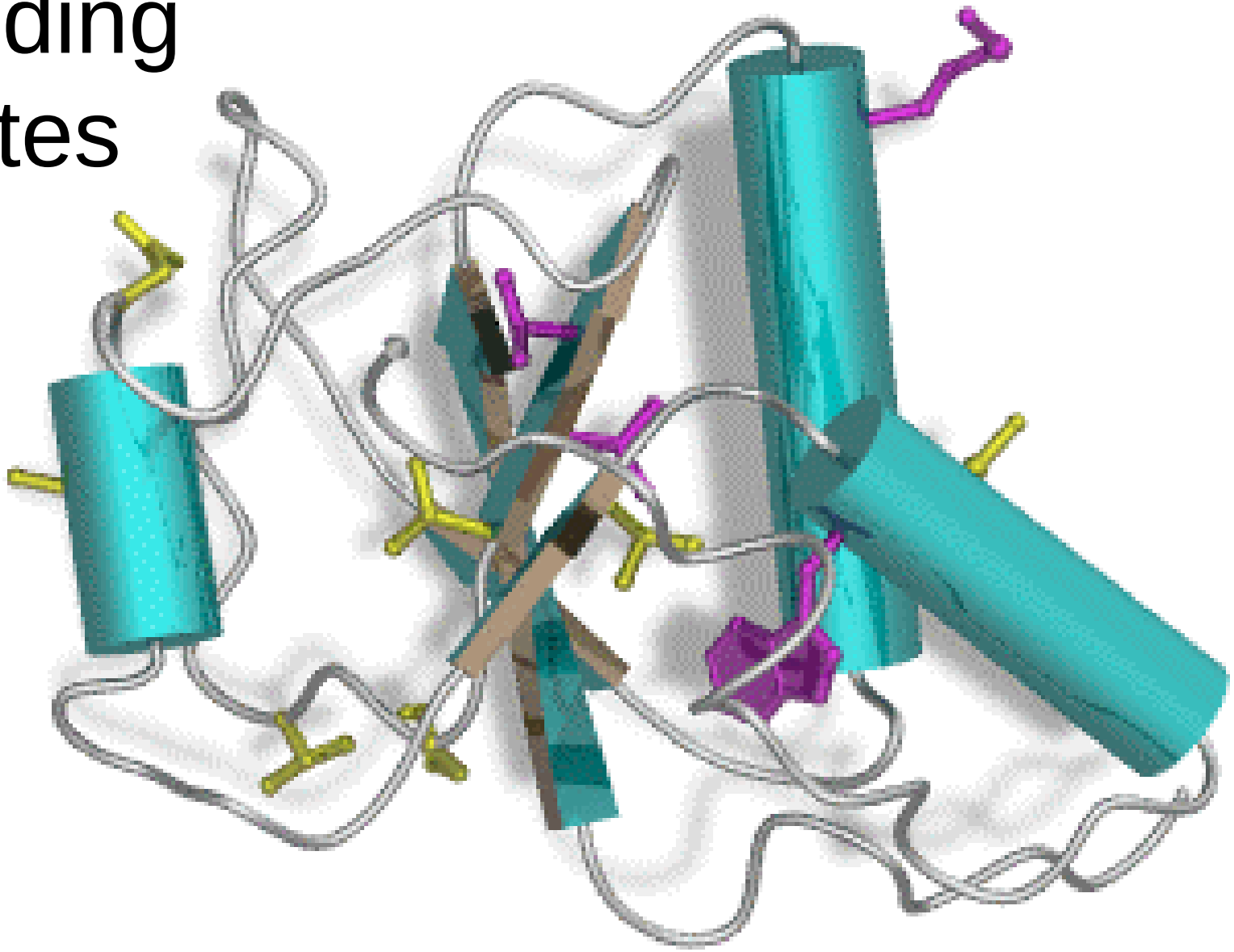
- Sequence → Structure → Function
- Approaches
  - biochemical (several years, phd)
  - simulation
    (molecular dynamics, *small molecules*)
  - prediction == search problem
    (heuristic methods / simplified models)
- Use: drug design, ...

Protein docking and ligand binding

- Protein docking:find the most stable mode of association between two protein molecules, starting from the atomic coordinates of the two isolated components.

- `lock and key' mechanism , where both lock and key are plastic, and distort according to mutual interactions.

Binding sites



Protein docking

- Aim - optimise the surface area and attractive forces and to minimise the loss of energy due to interaction with the solvent.
- Optimisation on many degrees of freedom, (6-D rigid body movement problem - 3 translations and 3 rotations, all of which must be searched)

Protein docking approaches

- Given the information of a pair of proteins *crystallised together*, **reconstruct** the docking
- Given the individual proteins *separately crystallised*, **predict** their docking. Requires trying all combinations of degrees of freedom
- Ligand binding - small ligands tend to bind in big pockets; ligands are more flexible than proteins

Next...

- Search and pattern discovery
  - sequences
  - structures
- Metabolic pathways
- Gene expression arrays
- Resources / reading

Search and pattern discovery

Sequences (DNA, RNA)
Structure (RNA, protein)

- Functionally significant regions, repeated in different entities, often described by patterns.
- Search through (very large) genome / protein databases for entries matching the pattern. (formal language theory,- [Searles93].
- Biological data is noisy: string languages - stochastic approaches
  - Hidden Markov models [Durbin et al, 98]
  - Stochastic context-free grammars [Lathrop&Smith, 96].

Sequence search & alignment

- Search for sequences of a few bases to kilobases long.
- Attempt to align unknown sequence against those on record.
- Gapped Sequences: deletion or insertion (result of a mutation)
- Various alignment programs
- Use of substitution matrices
- Filters: mask regions of query sequence with low compositional complexity
- 1998: >1200 million base pairs, >1.6 million sequences (EMBL)

Sequence alignment problem

```
T C A T G
 / / /  |
C A T T G
```

```
T C A T G
 / /  |  |
C A T T G
```

Alignment - Gaps

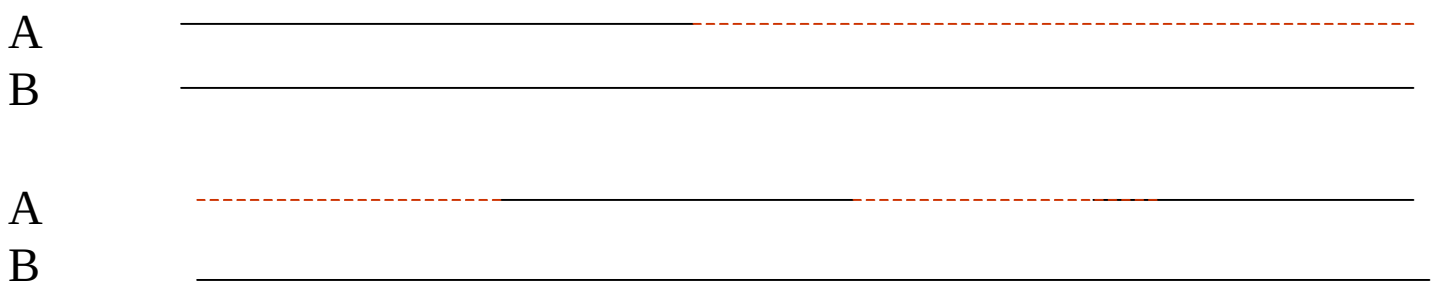
No gaps:

```
AGGVLIIQVG
||||||
AGGVLIQVG
Score = 6
```

With gaps

```
AGGVLIIQVG
|||||| |||
AGGVLI-QVG
Score = 9
```

… but add penalties for multiple gaps



Alignment via dynamic programming

- Needleman & Wunsch - global
- Smith-Waterman - local

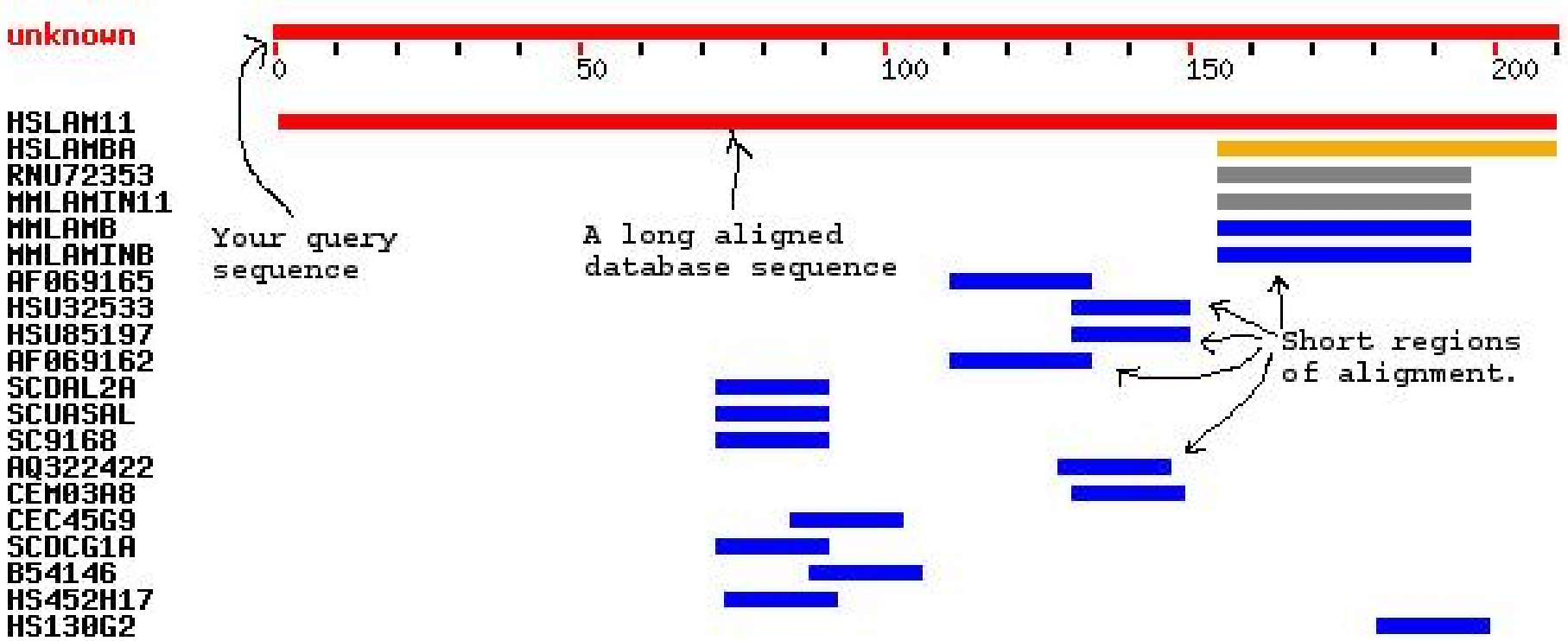
Pairwise database searching

- FAST A
- BLAST (Basic Local Alignment Search Tool), gapped-BLAST, PSI-BLAST

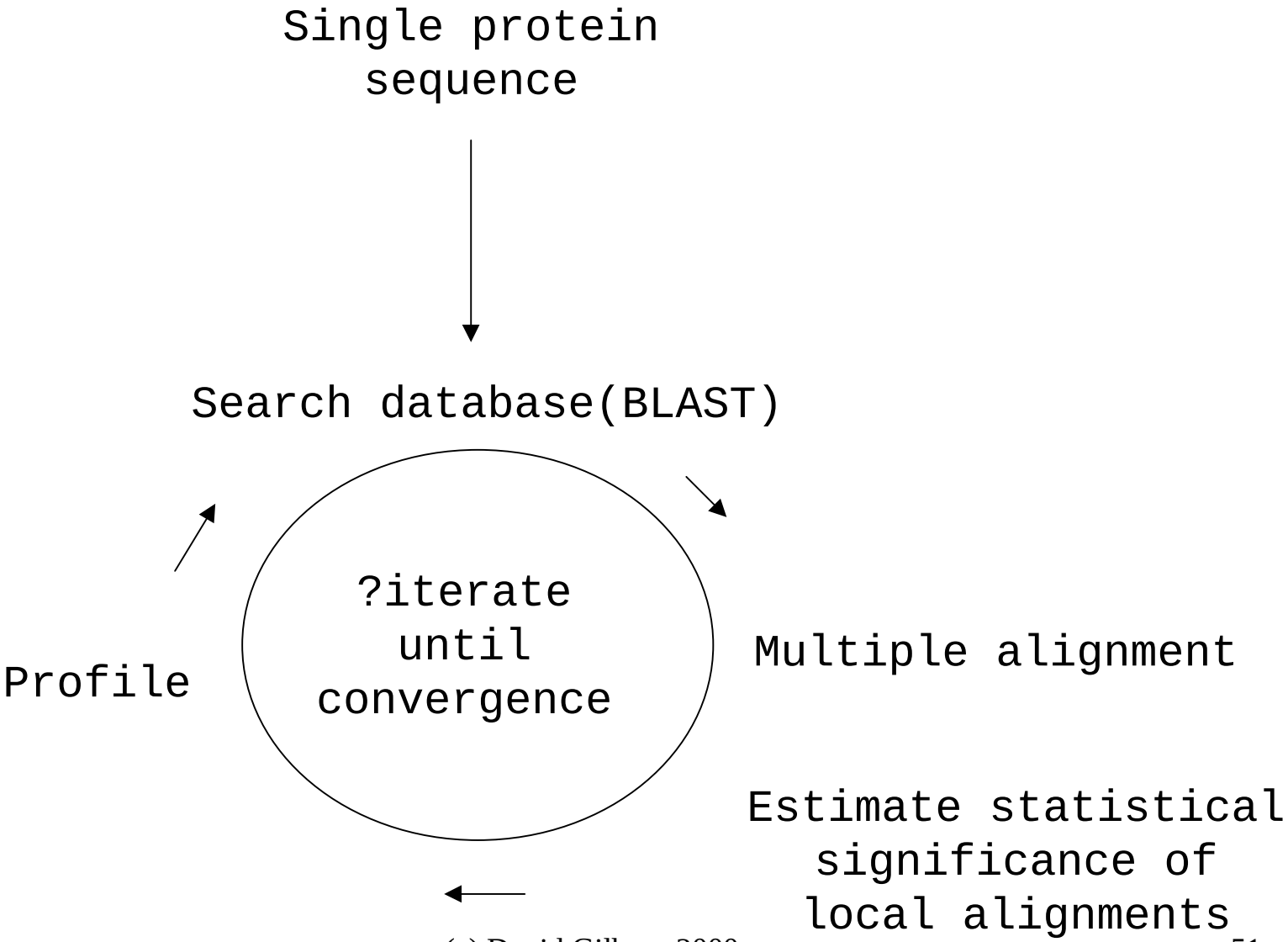
Human nucleotide sequence

AAAAGAAAAGGTTAGAAAGATGAGAGATGATAAAGGGTCCATTTGAGGTTAGGTAAT
ATGGTTTGGTATCCCTGTAGTTAAAAGTTTTTGTCTTATTTTAGAATACTGTGACTA
TTTCTTTAGTATTAATTTTTCCTTCTGTTTTCCTCATCTAGGGAACCCCAAGAGCAT
CCAATAGAAGCTGTGCAATTATGTAAAATTTTCAACTGTCTTCCTCAAAATAAAGAA
GTATGGTAATCTTTACCTGTATACAGTGCAGAGCCTTCTCAGAAGCACAGAATATTT
TTATATTTCCTTTATGTGAATTTTTAAGCTGCAAATCTGATGGCCTTAATTTCCTTT
TTGACACTGAAAGTTTTGTAAAAGAAATCATGTCCATACACTTTGTTGCAAGATGTG
AATTATTGACACTGAACTTAATAACTGTGTACTGTTCGGAAGGGGTTCCTCAAATTT
TTTGACTTTTTTTGTATGTGTGTTTTTTCTTTTTTTTTTAAGTTCTTATGAGGAGGGA
GGGTAAATAAACCACTGTGCGTCTTGGTGTAATTTGAAGATTGCCCCATCTAGACTA
GCAATCTCTTCATTATTCTCTGCTATATATAAAACGGTGCTGTGAGGGAGGGGAAAA
GCATTTTTCAATATATTGAACTTTTGTACTGAATTTTTTTGTAATAAGCAATCAAGG
TTATAATTTTTTTTAAAATAGAAATTTTGTAAGAAGGCAATATTAACCTAATCACCA
TGTAAGCACTCTGGATGATGGATTCCACAAAACTTGGTTTTATGGTTACTTCTTCTC
TTAGATTCTTAATTCATGAGGAGGGTGGGGGAGGGAGGTGGAGGGAGGGAAGGGTTT
CTCTATTAAAATGCATTCGTTGTGTTTTTTAAGATAGTGTAACTTGCTAAATTTCTT
ATGTGACATTAACAAATAAAAAAGCTCTTTTAATATTAGATAA

BLASTN results



PSI-BLAST (position specific iterated)



Pattern discovery in biosequences

- Motivation:
  - gene functional class prediction
  - RNA splicing
  - protein structure & function
  - gene regulation (transcription factor binding site prediction)

[Alvis Brazma & Inge Jonnassen]

Protein families

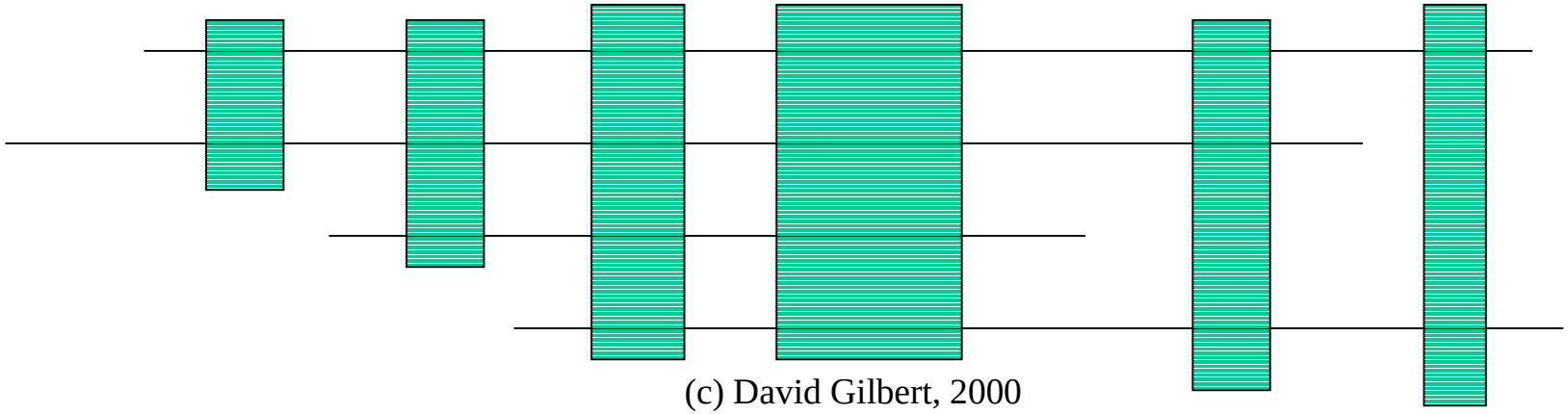
- Prediction of structure/function from sequence:
  - sequence database similarity search
  - compare to family descriptions
  - structure prediction programs

Protein family analysis

- Collect sequences (structures) in family
- Analyze
  - local multiple alignment
  - global multiple alignment
  - pattern discovery
- Make family description
- Pick up more family members?
  - Analyze extended set

Multiple vs. pairwise comparison

- Multiple sequence comparison
  - is more sensitive
  - provides more information
  - is more difficult

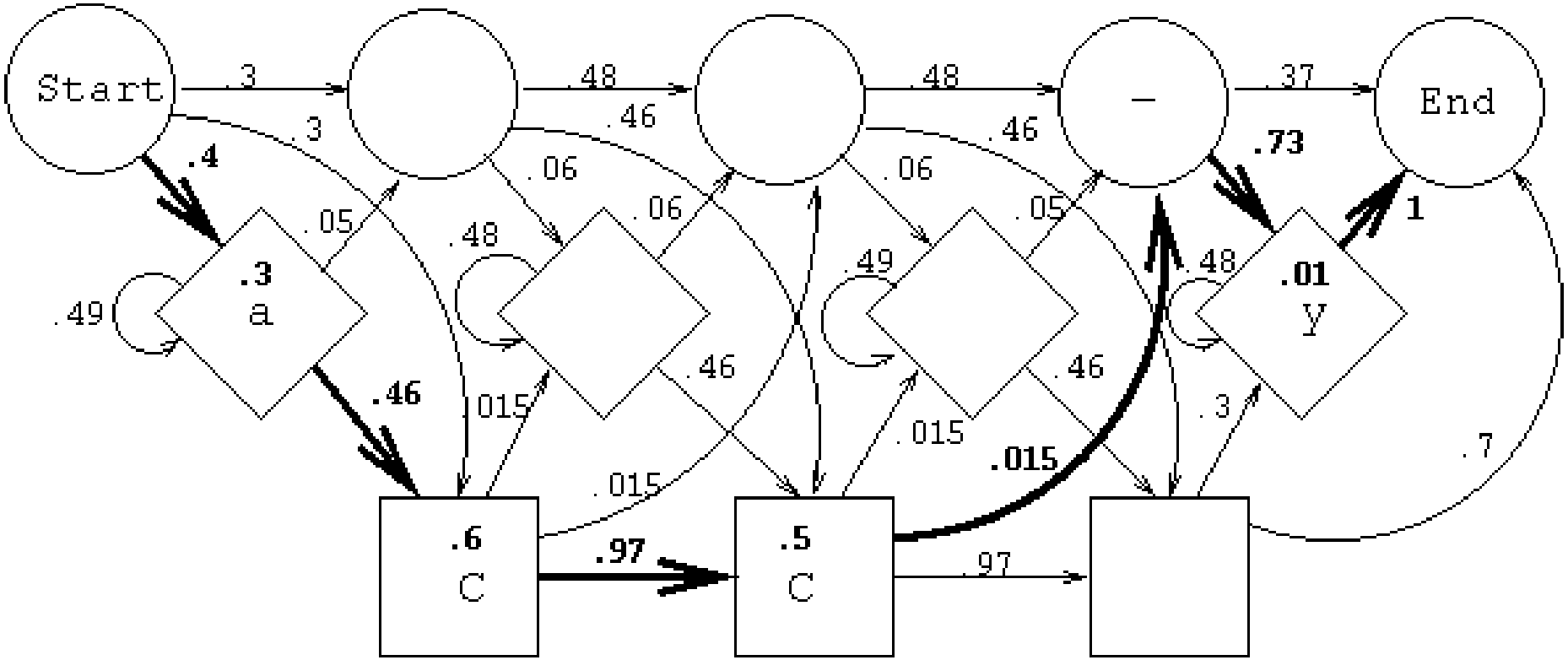


Patterns and alternative representations

- Patterns
  - unions of patterns
  - decision trees
  - exact/approximate matching
- Alignments, weight matrices, profiles, HMMs, Neural networks, SCFGA, ...

PROSITE profiles

- Uses Hidden Markov Model - can characterise an entire family of sequences.



Discrete patterns

- **Advantages**
  - simple and easily interpretable objects
  - easier to discover from scratch (i.e., if no additional information to sequences are given), particularly in noisy data
- **Disadvantages**
  - limited descriptive power (no weights can be attributed to alternatives)

Biosequences - general

- Basic alphabet
  Σ = { a, t/u, c, g} (DNA/RNA)
  Σ = {A, C, .., Y} (Protein sequence)
- Character group alphabet $\Pi = \{g_1 \ldots g_n\}$
  (e.g. amino-acid class)
- Wild card X = $\{ x(n_1, n_2) \mid n_1 < n_2 \in N\}$
- $V(x(c_1,c_2))$ set of all words over Σ of length between c_1 and c_2)
- Pattern P = $p_1 \ldots p_n$, $p_i \in \Sigma \cup \Pi \cup X$

$\rightarrow$ character & position constraints $\leftarrow$

PROSITE

- Database of protein families and domains
- Consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs

PROSITE patterns

- `**x**' any amino acid
- Ambiguities :

  [ALT] =Ala or Leu or Thr

  {AM} any amino acid except Ala and Met.
- `-' separator, `<` N-terminal, `>` C-terminal
- `.` end of pattern
- Repetition: x(3) = x-x-x
- x(2,4) = x-x *or* x-x-x *or* x-x-x-x.

PROSITE examples

- [AC]-x-V-x(4)-{ED}.
  - [Ala or Cys]-x-Val-x-x-x-x-{any but Glu or Asp}

- <A-x-[ST](2)-x(0,1)-V.
  - Start at N-terminal of the sequence
  - Ala-x-[Ser or Thr]-[Ser or Thr]-(x or none)-Val

How to obtain these patterns?

Learning

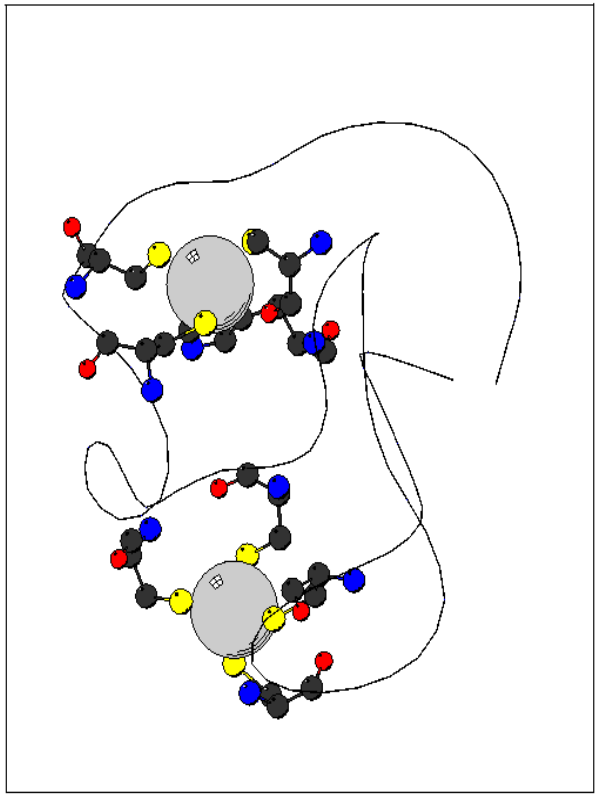
- Automatically find pattern (given a training set)

- Characterisation: (positive examples only) patterns describing “interesting” properties of a family

- Classification: (positive **and** negative examples) pattern distinguishing S+ and S- .. Which may overlap...

Example family (zinc finger c2h2)

```
      xxx
     V   x
    x     x
    x     x
    x     x
     C   H
    x \ / x
   x   Zn  x
    x /  \ x
     C    H
xxxx       xxxxxx
```



C-x(2,4)-C-x(3)-[ILVMFYWC]-x(8)-H-x(3,5)-H

Example property

A given sequence belongs to the chromo-domain family if it matches either the pattern:

E-x(0,1)-E-E-[FY]-x-V-E-K-[IV]-[IL]-D-[KR]-R-x(3,4)-G-x-V-x-Y-x-L-K-W-K-G-[FY]-x-[ED]-x-[HED]-N-T-W-E-P-x(2)-N-x-[ED]-C-x-[ED]-L-[IL]

or the pattern:

L-x(2,3)-E-[KR]-I-[IL]-G-A-[TS]-D-[TSN]-x-G-[EDR]-L-x-F-L-x(2)-[FW]-[KE]-x(2)-D-x-A-[ED]-x-V-x-[AS]-x(2)-A-x(2)-K-x-P-x(2)-[IV]-I-x-F-Y-E

or the pattern:

Y-x(0,2)-L-[IV]-K-W-x(6)-[HE]-x-[TS]-W-E-x(4)-[IL]

Various ways of using pattern matching for family characterization

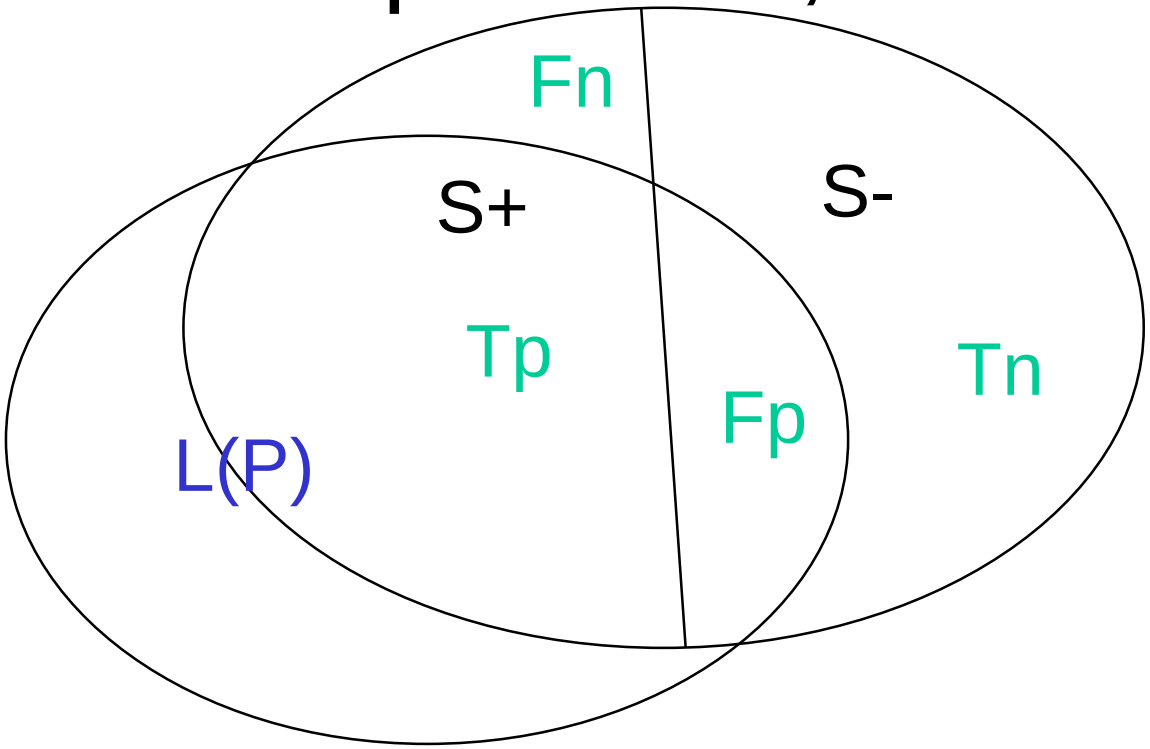
A sequence belongs to the family if

1. it matches the given sequence *pattern*;
2. if it is within a certain *distance* from a string that matches a the pattern (distance between strings can be defined either as a number of mismatches, or as an edit-distance, or based on similarity matrices or some other way) ;
3. if it matches one of a given set of patterns (i.e.,if it matches a union of patterns);
4. if a decision-tree over the matching patterns returns "yes"

Clean / Noisy Data

- Clean data: the training set is assumed to be "correct"
- Noisy data: training set
  - sequences may contain errors
  - sequences may have been assigned to the wrong family

True positives, true negatives, false positives, false negatives



Tp - true pos
Tn - true neg
Fp - false pos
Fn - false neg

$Tp = L(P) \cap S+$

$Tn = \neg L(P) \cap S-$

$Fp = L(P) \cap S-$

$Fn = \neg L(P) \cap S+$

L(P) - the set of sequences matched by the pattern P

Approaches to pattern discovery

- **Pattern driven:**

  enumerate all (or some) patterns up to certain complexity (length), for each calculate the score, and report the best

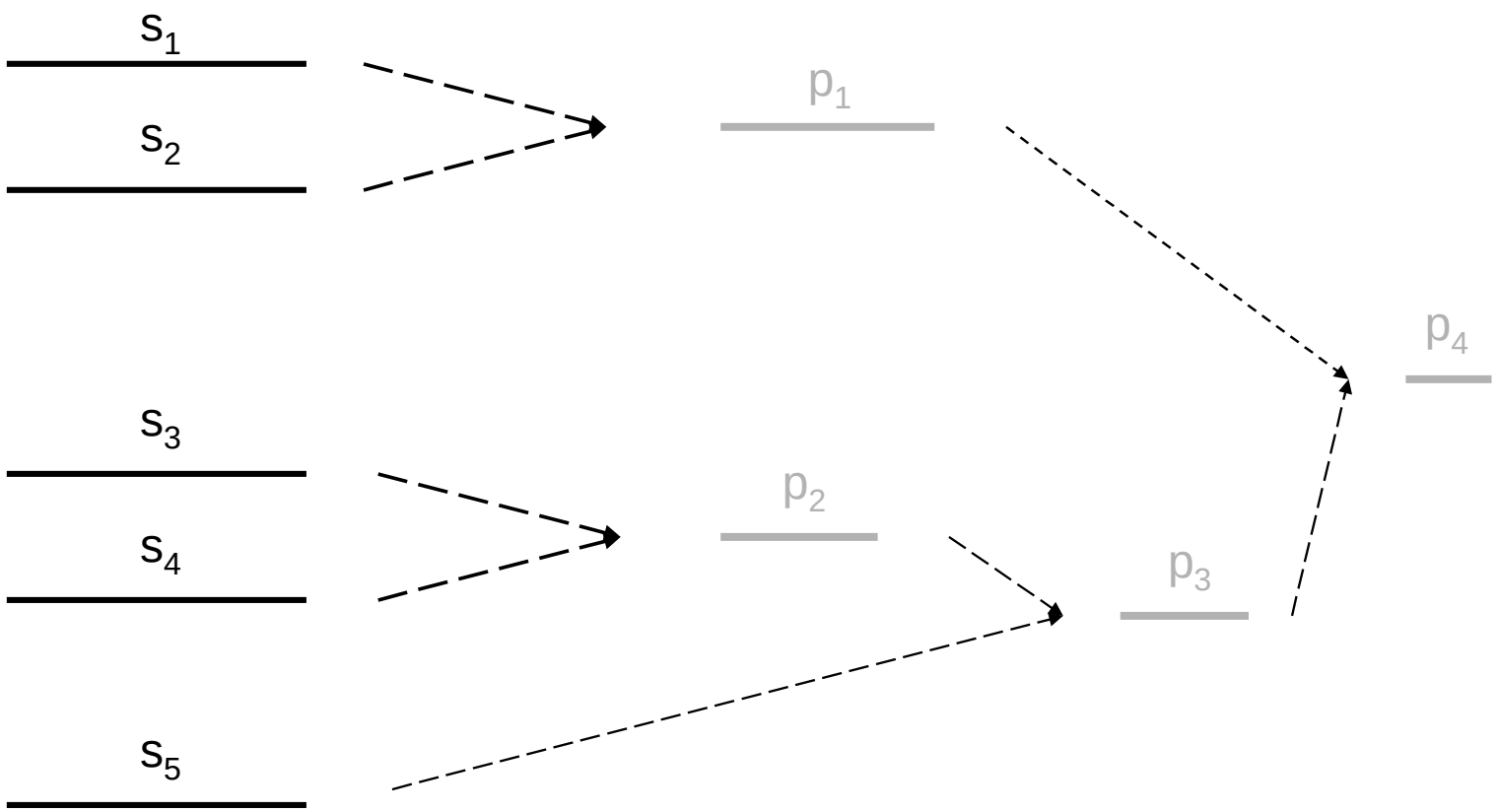
- **Sequence driven:**

  look for patterns by aligning the given sequences

Pattern driven algorithms

- Brute force - enumerate all patterns (for instance, all substrings) up to a given length (complexity)
- Evaluate their fitness with respect to the input sequences and output the best
- Unrealistic for patterns of even modest size even for substring patterns (e.g., for substring patterns of length 10 over the amino acid alphabet, there are more than 10^{13} different substrings to enumerate in this way)

Sequence driven approach



Sequence driven algorithms

- Group similar sequences together (e.g., in pairs);
- For each group find a common pattern (e.g., by dynamic programming);
- Group similar patterns together and repeat the previous step until there is only one group left

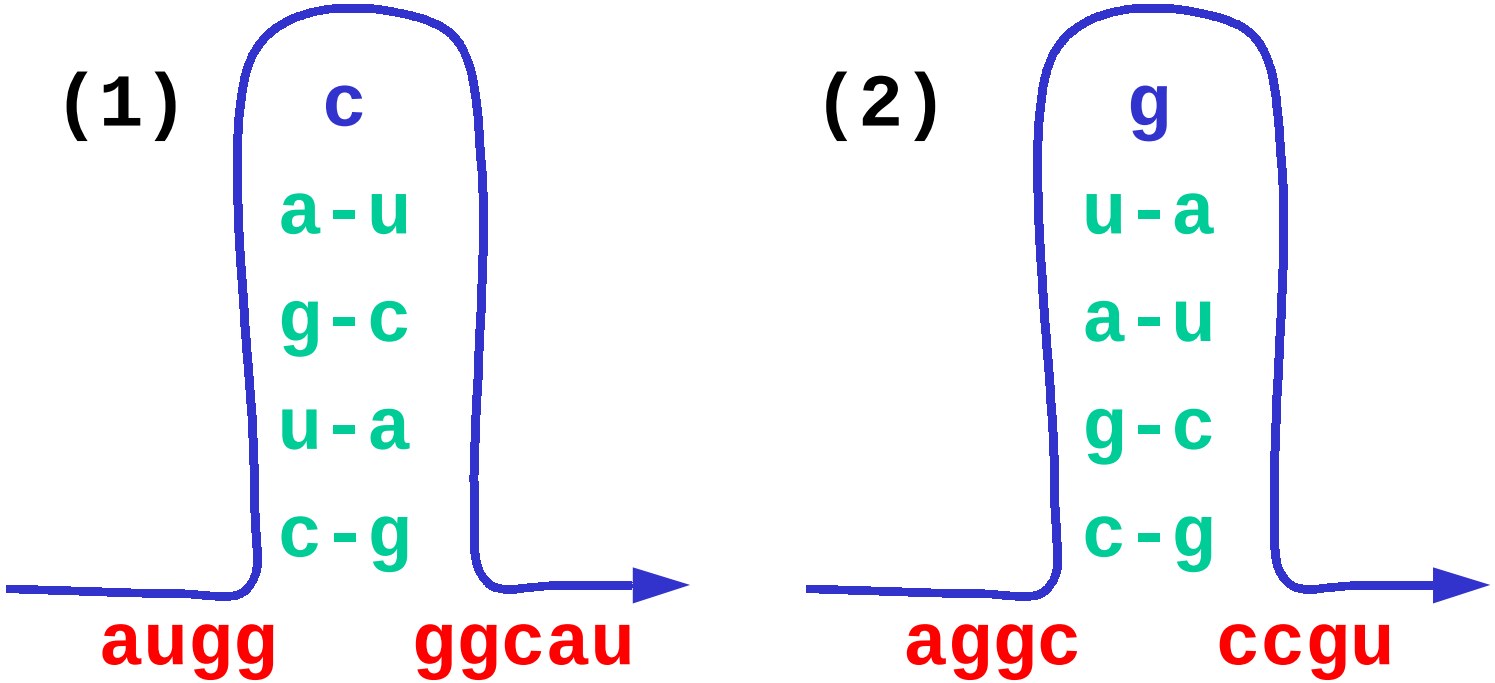
RNA structural patterns

- Constraints:
  - string length
  - inter-string distance
  - character contents
  - matching positions
  - correlation (identical, reverse, complement).
- Complements **`a-u g-c, g-u`** (weaker)
- Structures: Stem-loops, Pseudo-knots, Clover leafs
- CFG

Possible patterns

- Tandem repeat α-α acg acg
- Simple repeat α-β-α acgaaaacg
- Multiple repeat α-β-α-δ-α
  acgaaacguuacg
- Stem loop α-β-α^{rc} acgaacgu
- Palindrome α-α^{rc} acg gca
- Pseudoknot α-γ_1-β-γ_2-α^{rc}-γ_3-β^{rc}
  auggcugaaggccgaucucagggcauaucgccgu

Stem loops

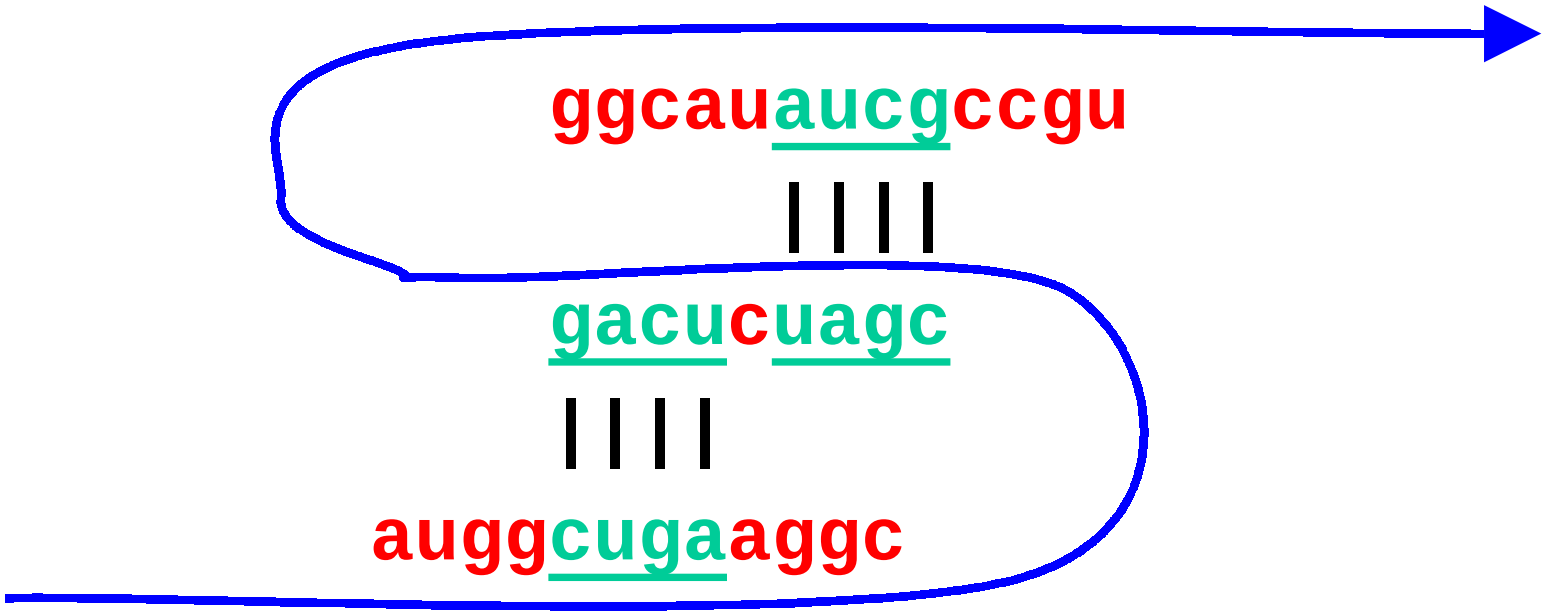


(1) auggcugacucagggcau

(2) aggccgaugaucgccgu

$\alpha \quad \beta \quad \alpha^{rc}$

Pseudo-knot

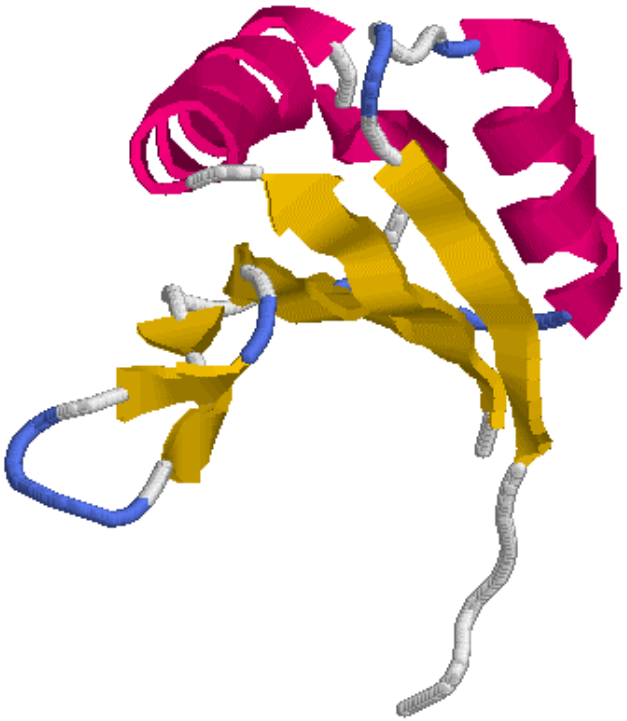


String:

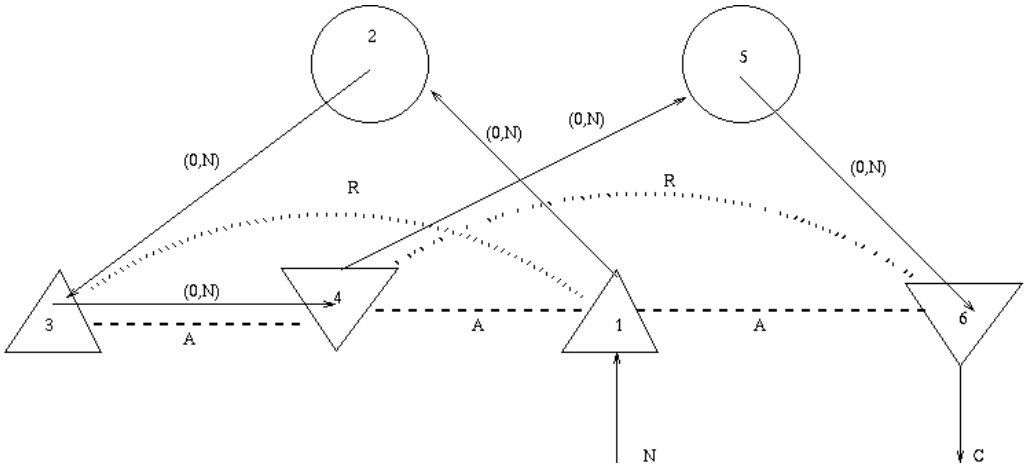
auggcugaaggccgaucucagggcauaucgccgu

$\alpha \quad \gamma_1 \quad \beta \quad \gamma_2 \quad \alpha^{rc} \quad \gamma_3 \quad \beta^{rc}$

Protein patterns (motifs)



Instance
(2bop)



Plait motif

Topological pattern discovery (Pattern-driven)

- Repeat:
  - Find new sheet
  - Extend sheet (linear)
  - Find circuits
- Works (in theory) on set of any size

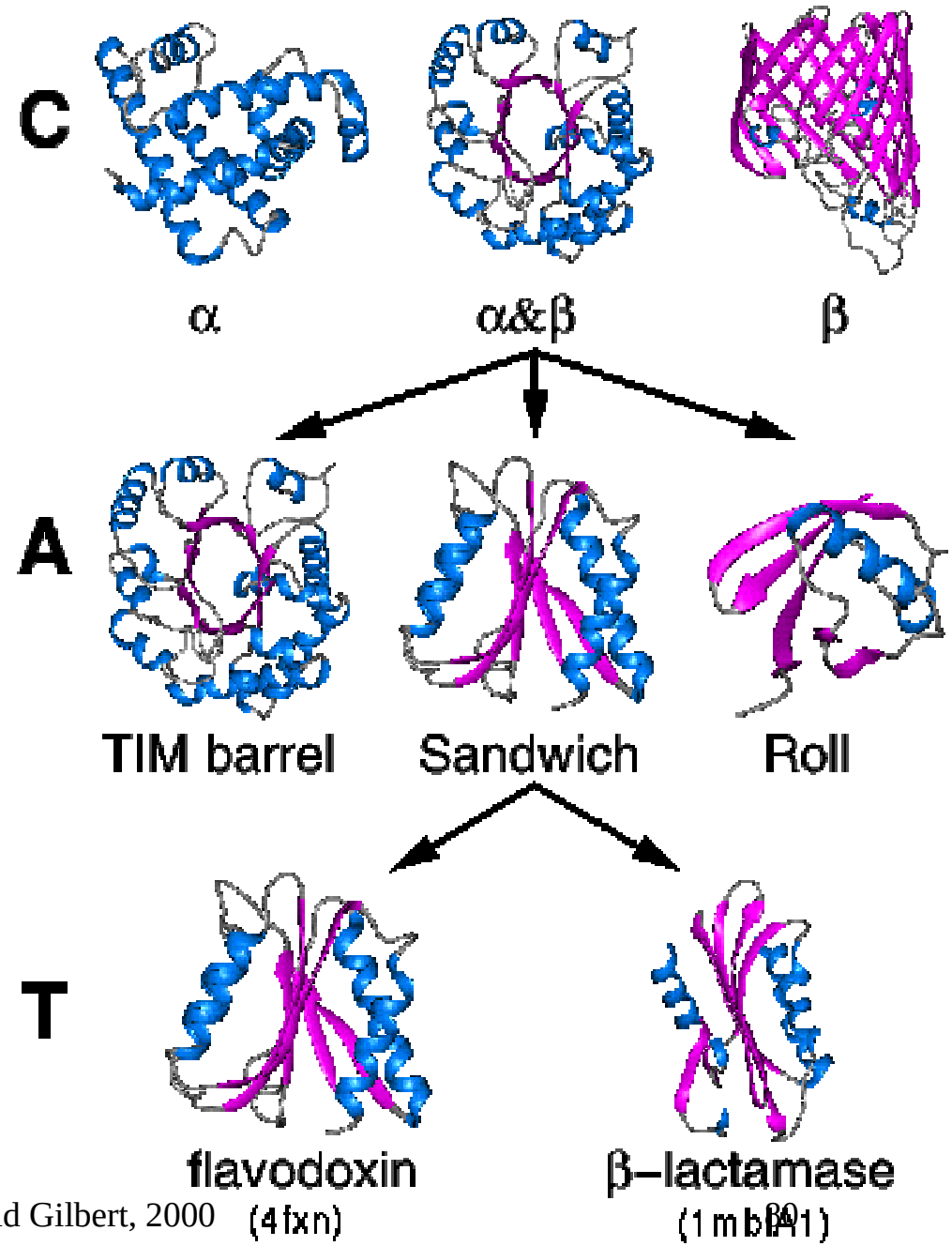
Based on pattern extension and repeated matching

Topological pattern discovery - maximal cliques

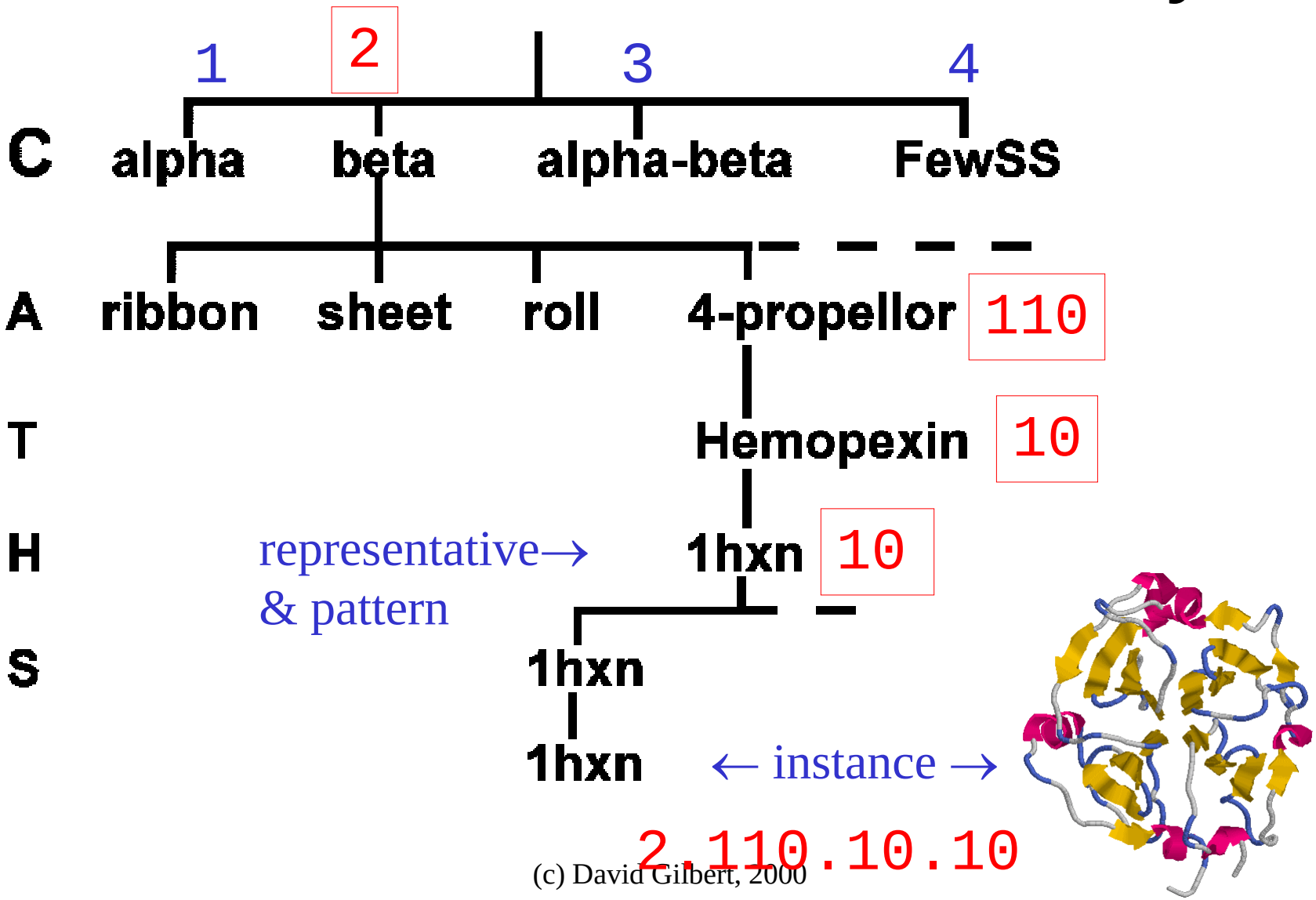
- Maximal clique detection in edge product graphs, (Bron-Kerbosch algorithm)

- only practical for pairwise comparisons

CATH hierarchy



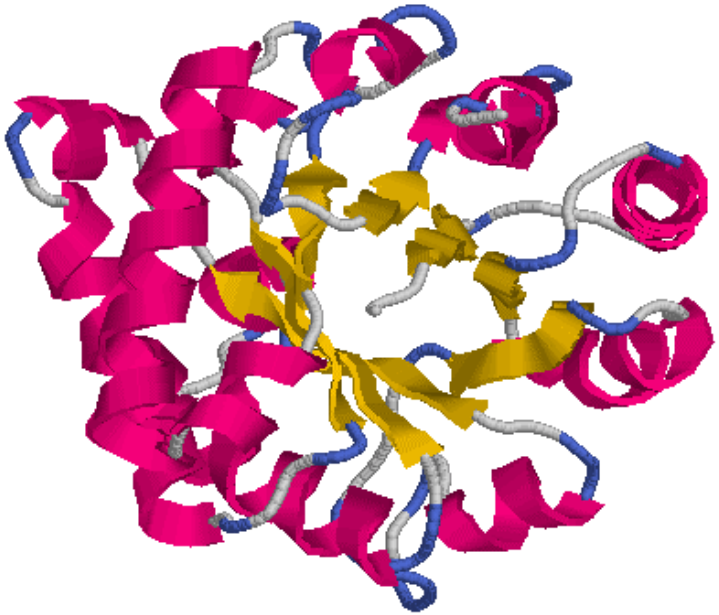
CATH database hierarchy

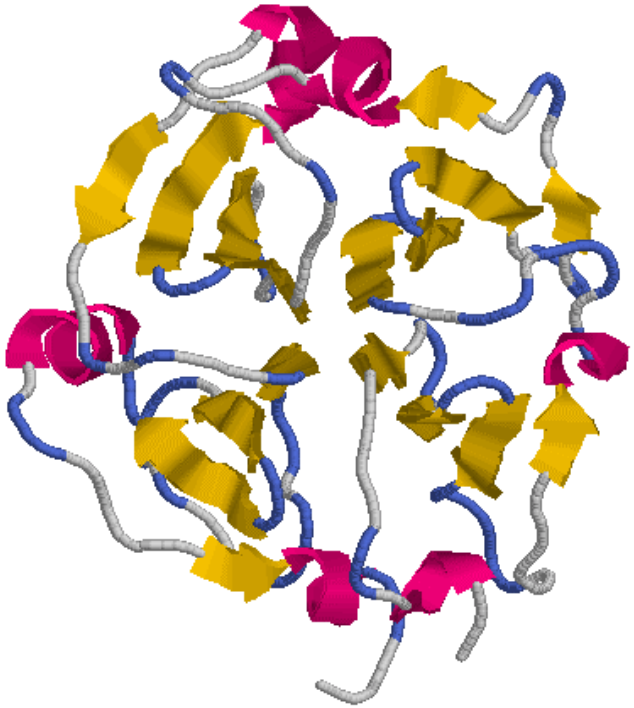


Structure comparison

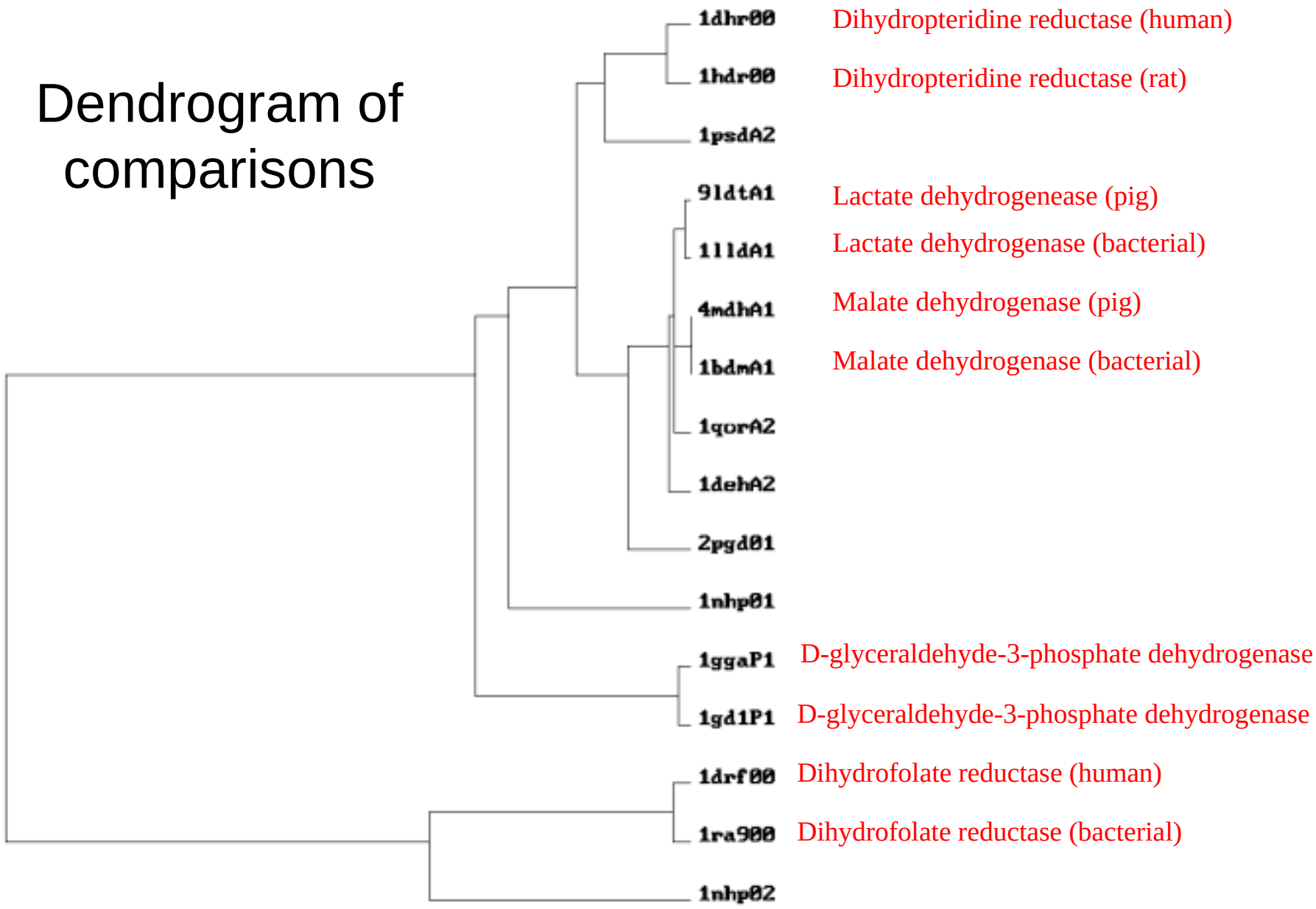
- Atomic coordinate level (RMSD)
- Threading and double dynamic programming
- Graph comparison
- Alignment using discovered patterns
- Issues:
  - validation (Gold Standard = Alexei Murzin)
  - distance metric (triangle inequality)

Structure comparison





Dendrogram of comparisons

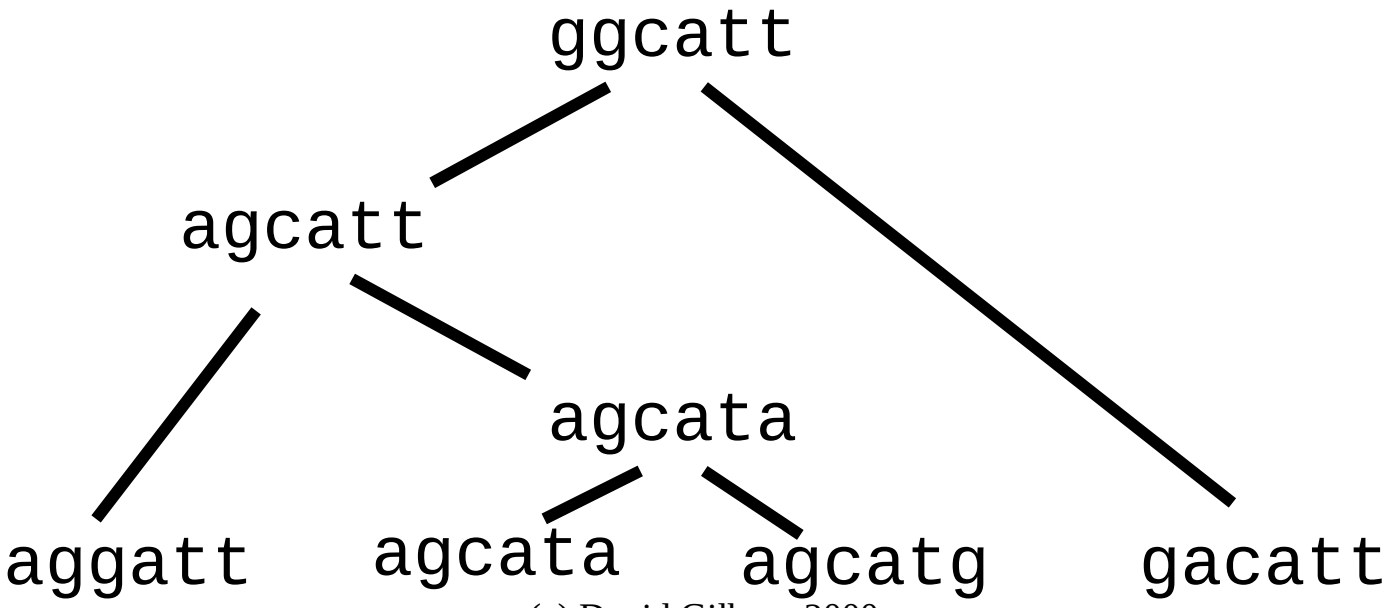


Phylogenetic Trees

- Phylogeny: relationship between species
- Phylogenetic tree: visualising evolution from a common ancestor.
- Labels - distance measure (e.g. time when the species evolved from a common ancestor.
- Gene divergence:
  - speciation = orthologues
  - duplication = paralogues

Evolution

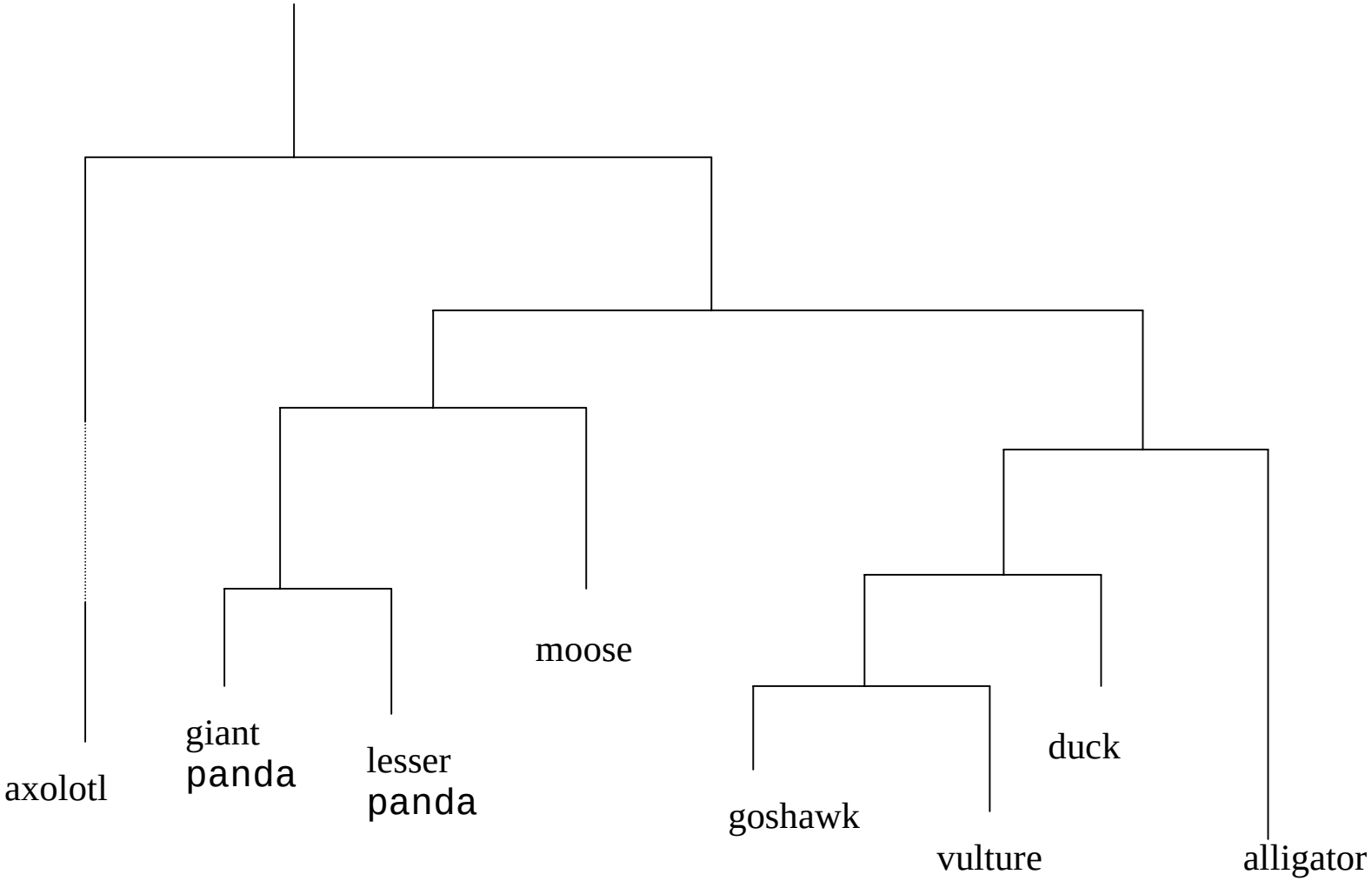
- proceeds primarily by duplication of genes, followed by divergence of function through mutation
- bioinformatics - detect distant similarities (homologies) in present-day sequences
- sheds light on evolution



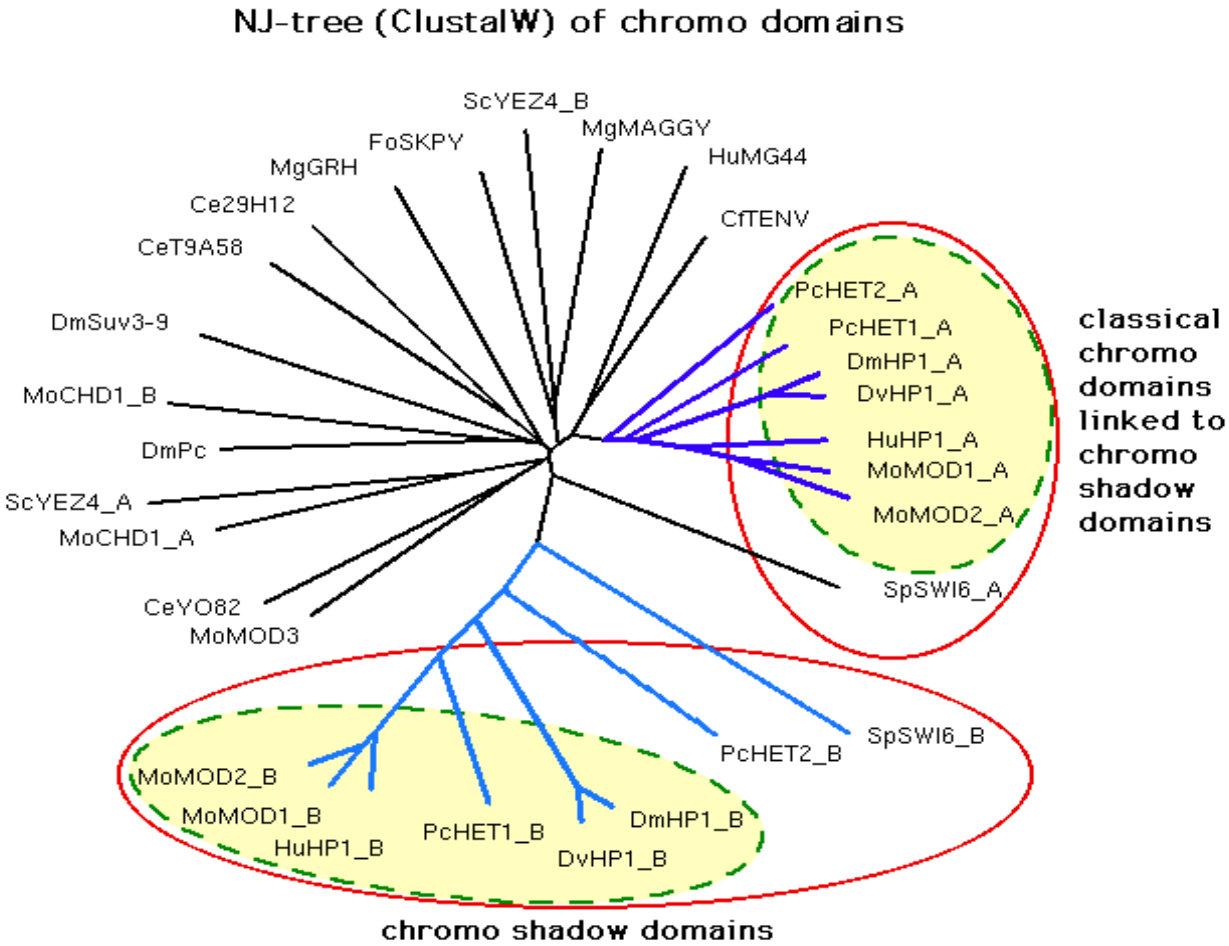
Phylogenetic Trees

- Clustering
- Constructed from pairwise distances by a variety of methods,
  - UPGMA (unweighted pair group method using arithmetic averages) [Sokal & Michener1958]
  - Parsimony e.g [Fitch1971].
- Bootstrap method [Feldenstein1985] used give measure of confidence for the tree.

Orthologue α-haemoglobins



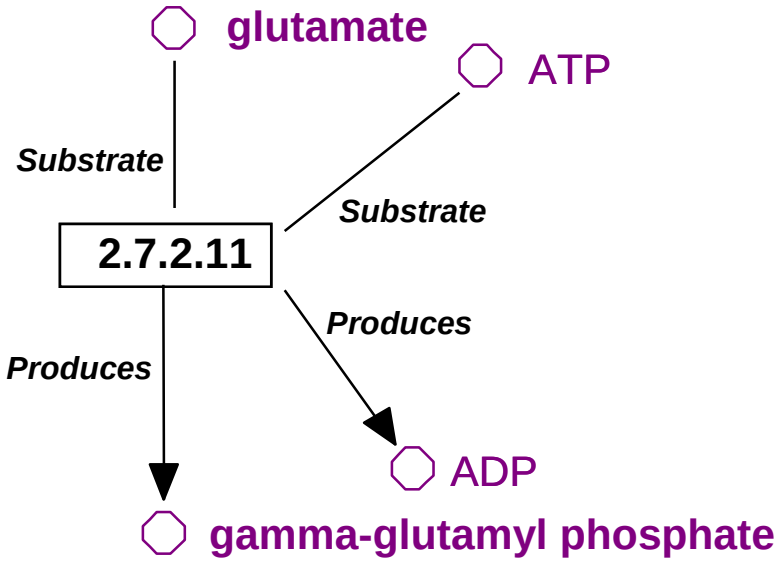
Unrooted tree



Biochemical Networks

- *metabolic reactions* transform *substrates* into *products*
- *metabolic pathways* - chains of reactions
- reactions occur spontaneously at an extremely slow rate.
- each reaction *catalyzed* by specialised proteins - *enzymes*.
- *enzymes* regulated by controlling either their *level of expression* or their *activity*.

Chemical Reaction



List of Biochemical Entities (substrates)

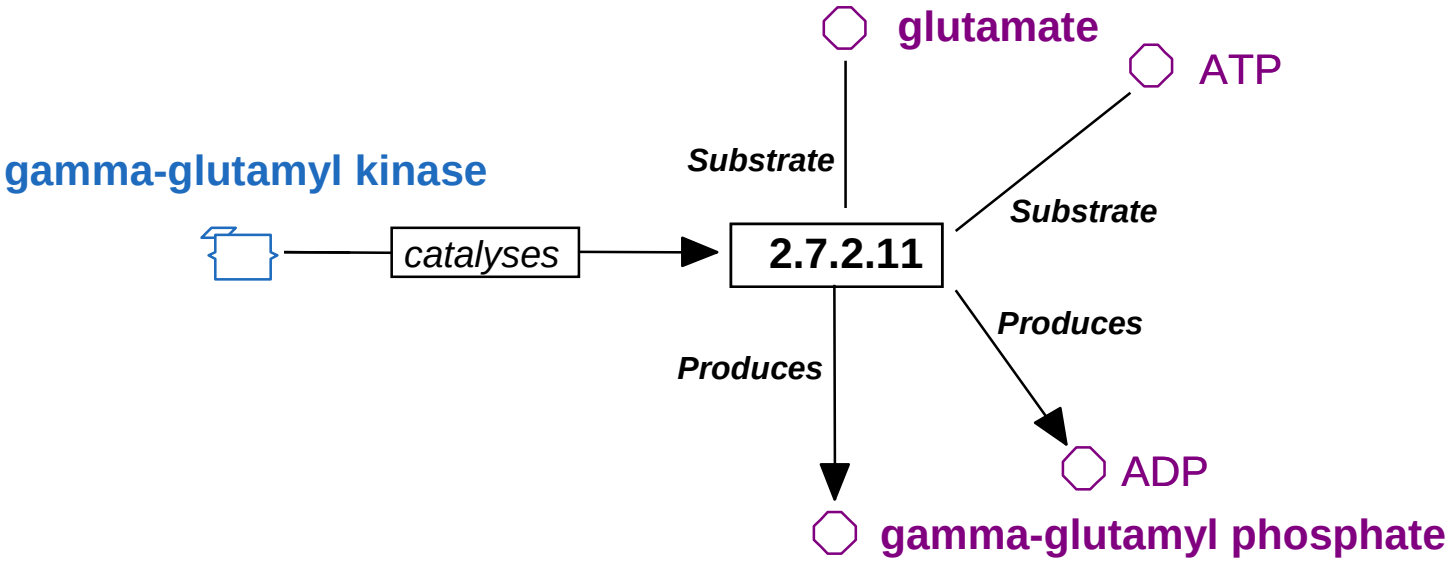
-o [Reaction] ->

List of Biochemical Entities (products)

1.5.1.2 **EC (reaction) number**

compound

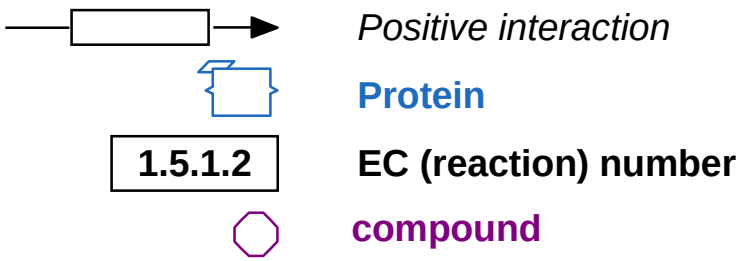
Enzymatic catalysis



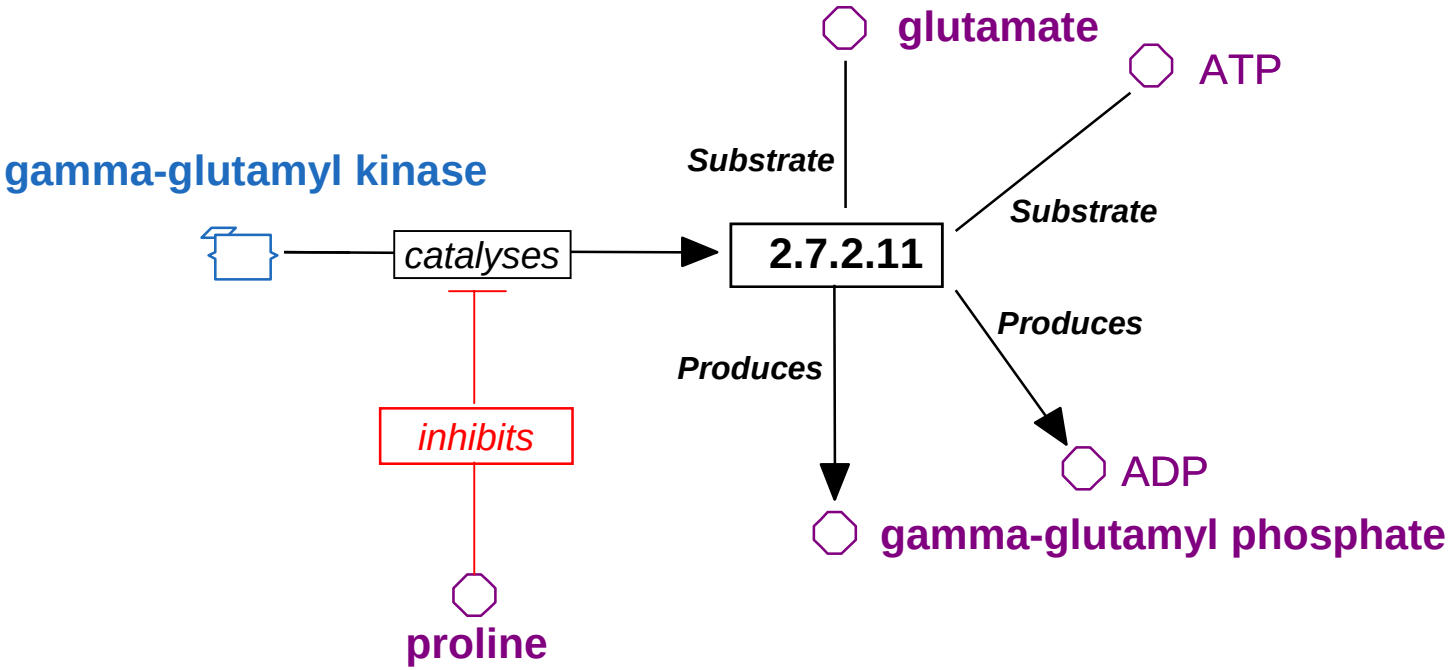
Protein (enzyme)

-o [Catalyses] ->

Reaction



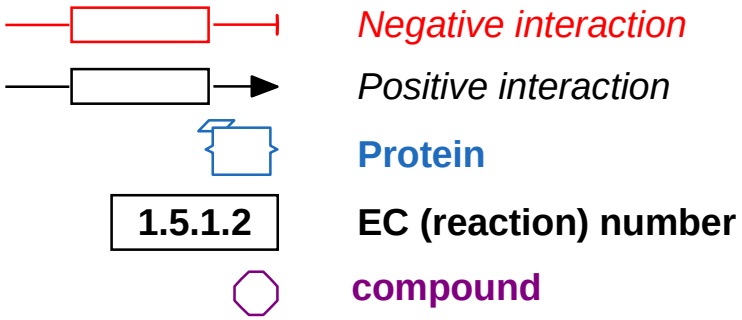
Inhibition/Activation



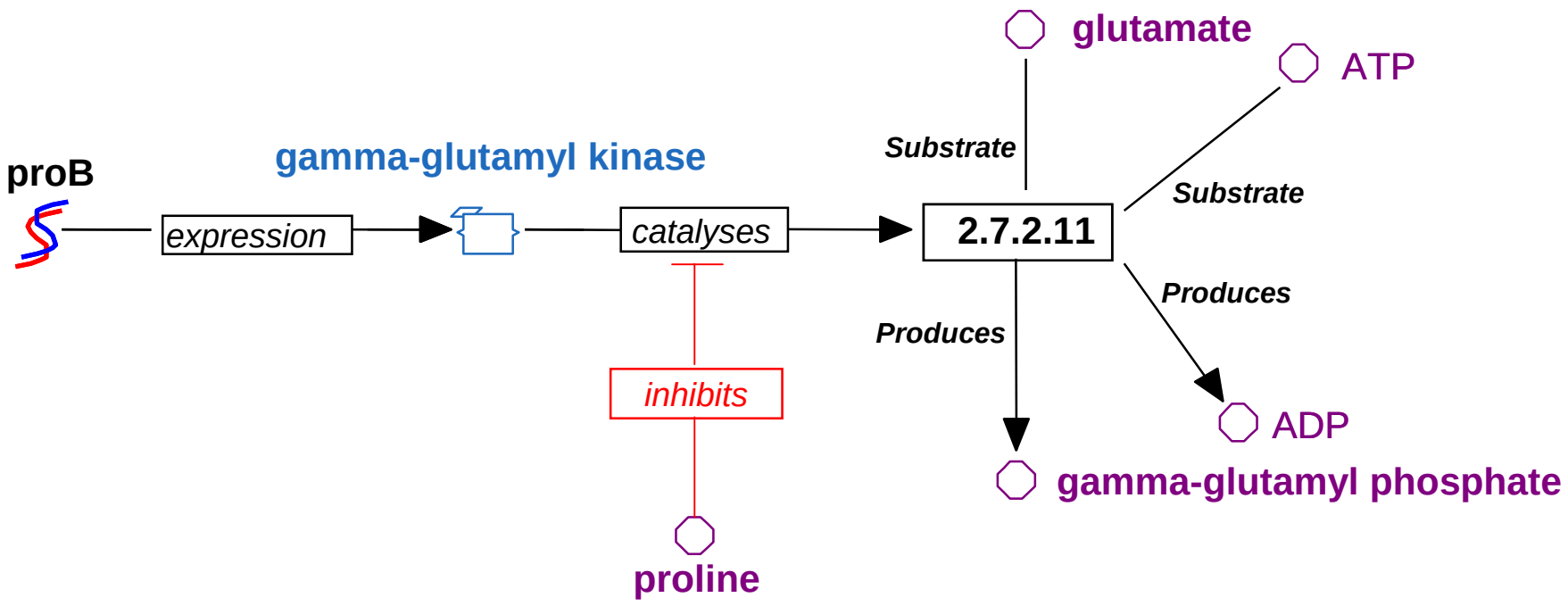
Biochemical Entity

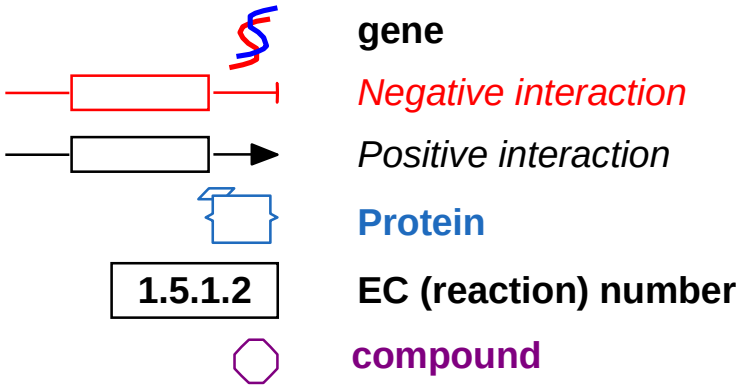
-o [Inhibits] ->

Reaction Catalysis



Metabolic Step



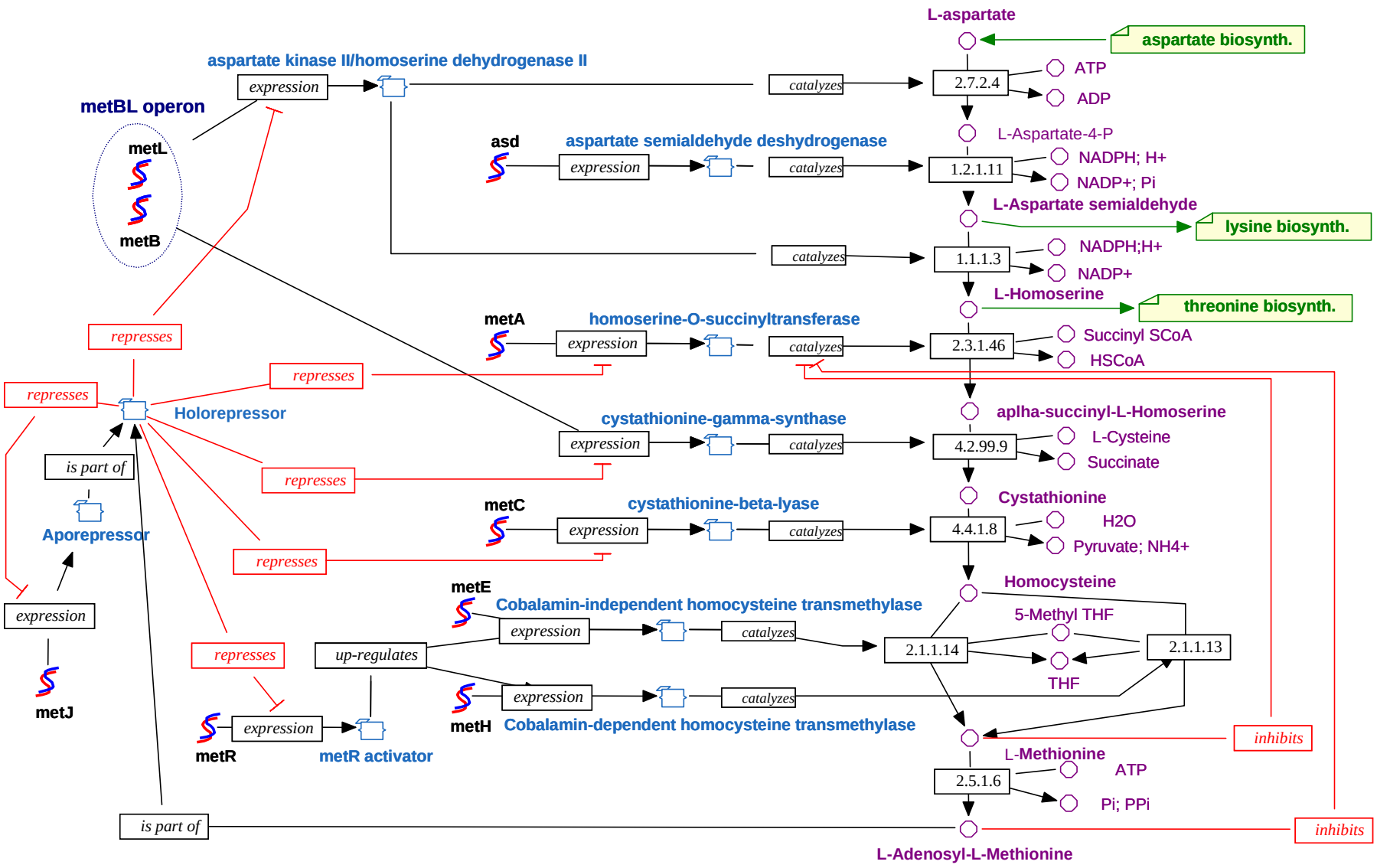


Biochemical Entity

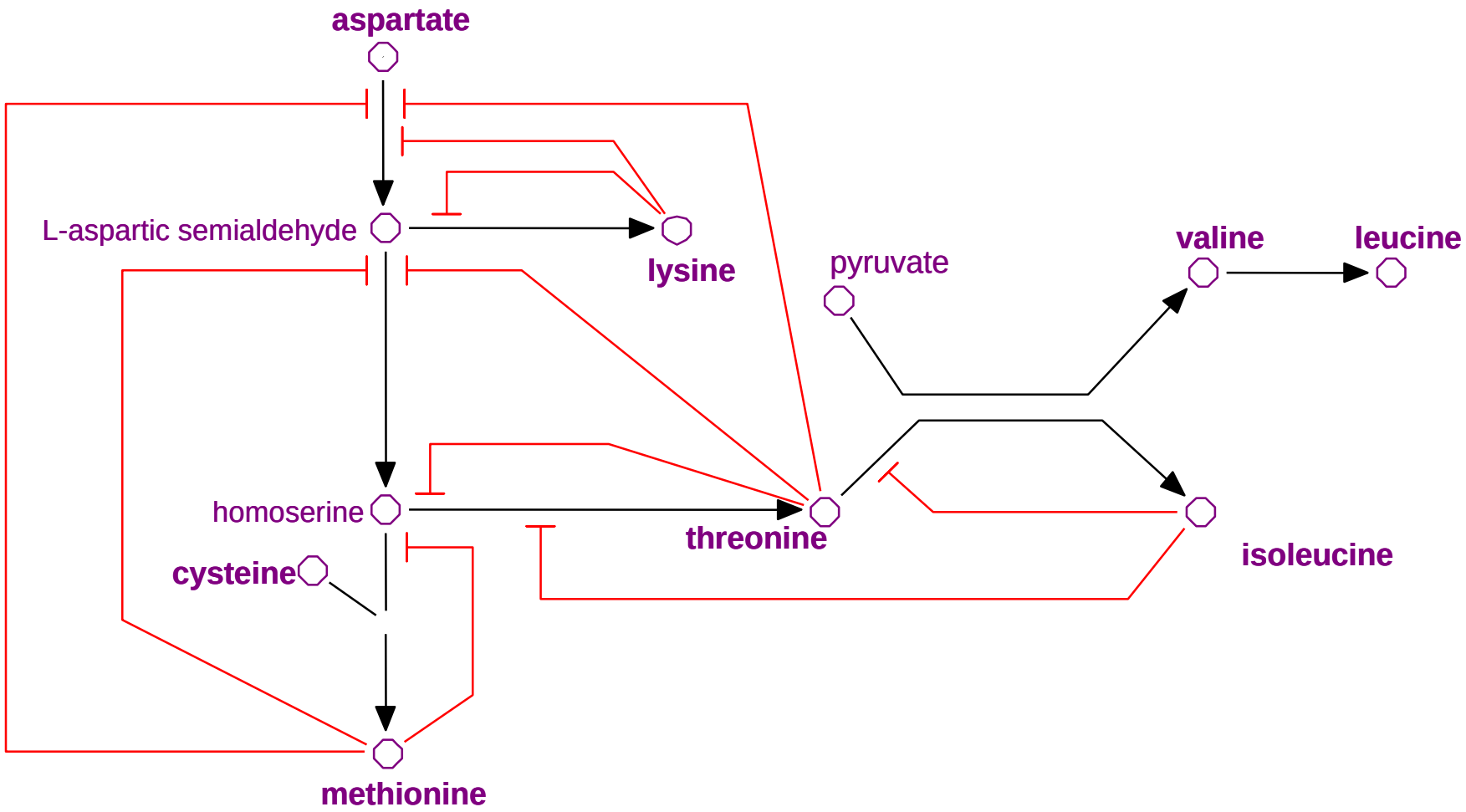
-o [Inhibits] ->

Reaction Catalysis

Methionine Biosynthesis in E.coli



High-level Abstraction



Networks - qualitative computation

Network navigation:

- How many pathways/ steps within each pathway, from A to compound B
- Give all the pathways that contain / lack specified compounds or processes
- Highlight pathways/networks according to various
- When genes/proteins are turned off or missing, show which paths or pathways may be affected.

Network analysis

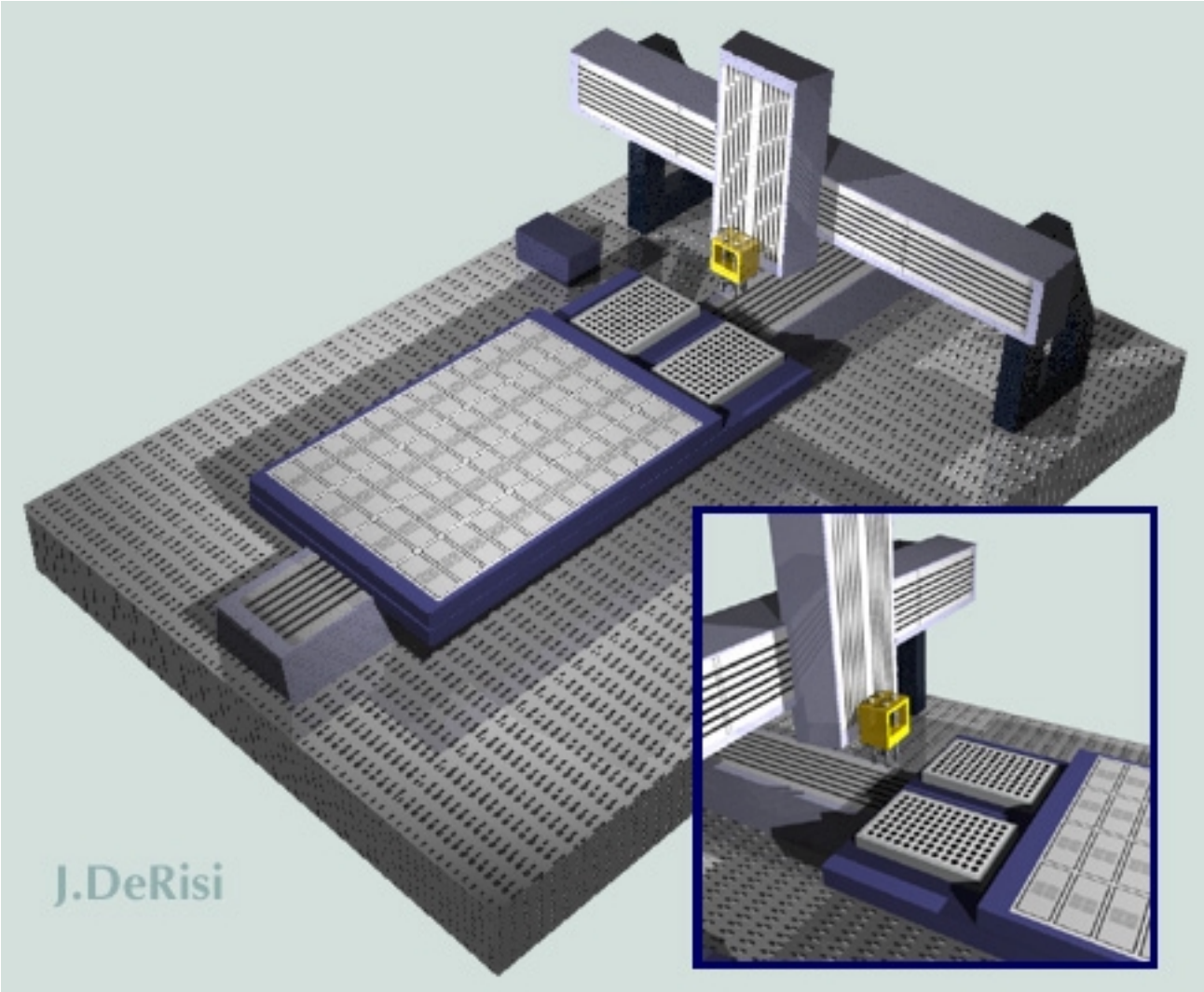
- Compare biochemical pathways from different organisms and tissues, or at different stages of annotation; highlight common features and differences; predict missing elements ('reconstruction')
- Represent pathways at different resolution
- Compile repertoires of recurrent network motifs at different resolution levels
- Identify all positive/negative regulatory cycles in a pathway graph.

Computation over networks

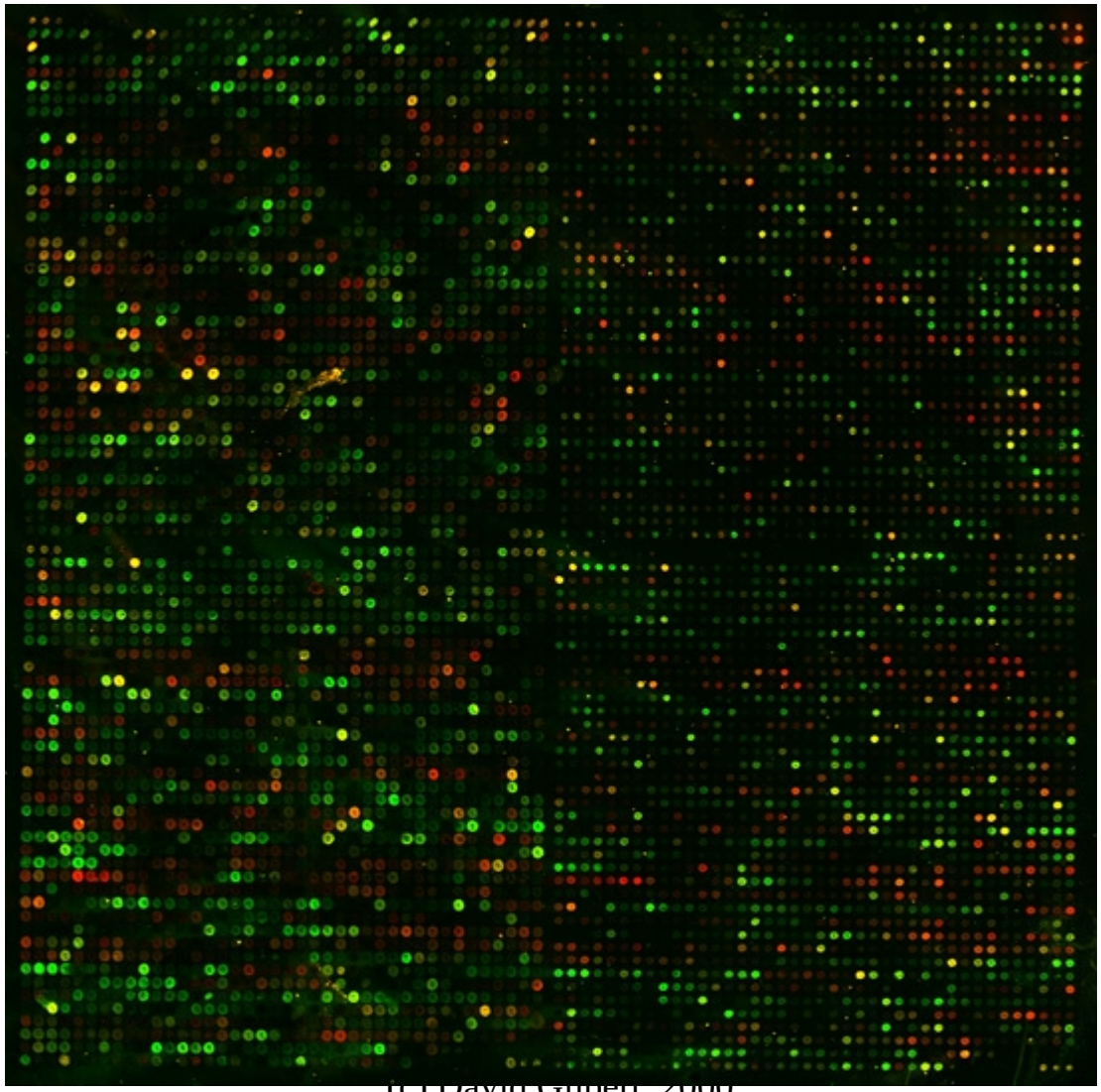
- Quantitative - Simulation
- Qualitative - Analysis
  - graph based
  - pi-calculus
- Both - e.g. petri nets
- Display - automatic graph layout

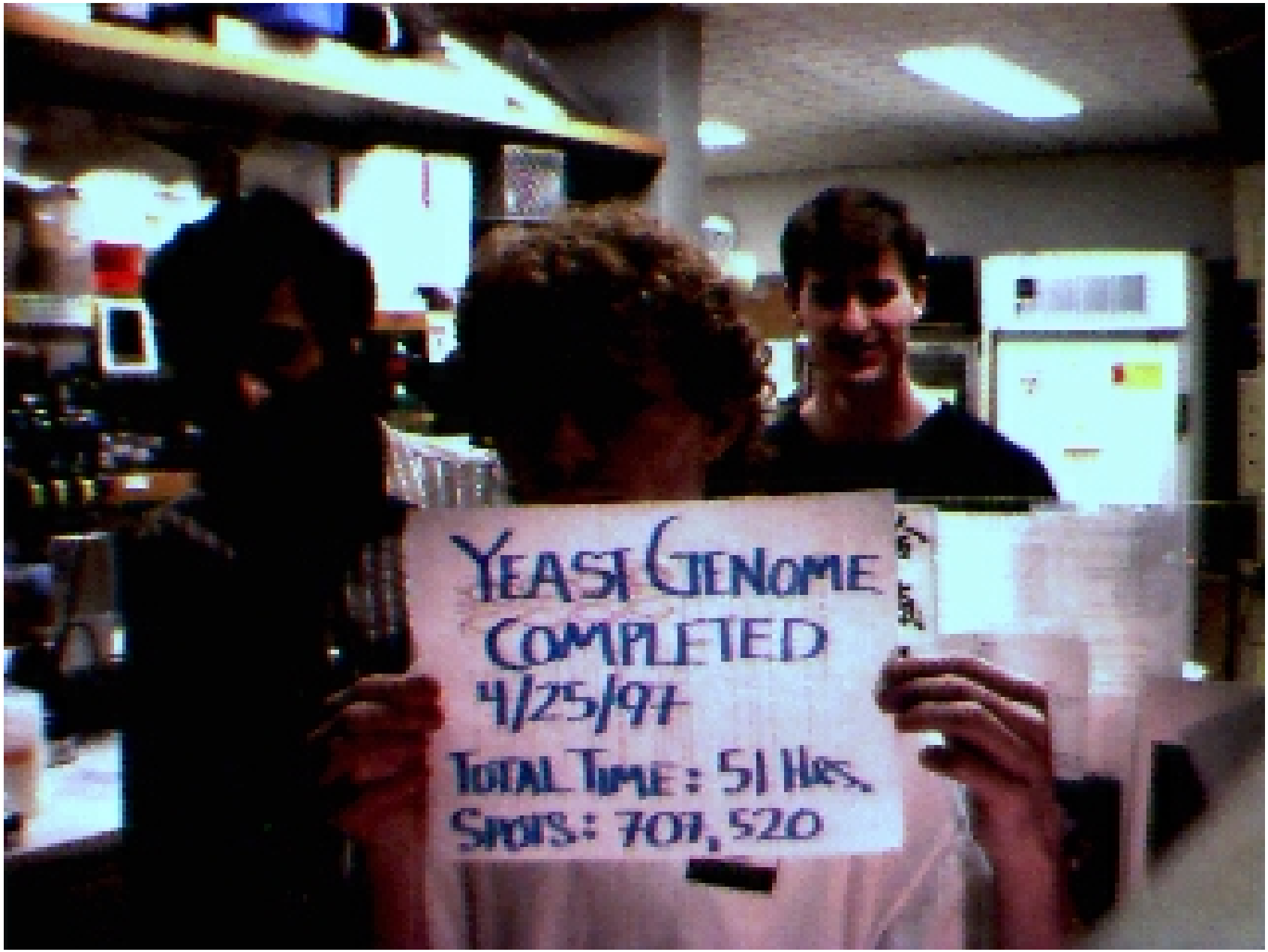
?Constraints?

Array technology

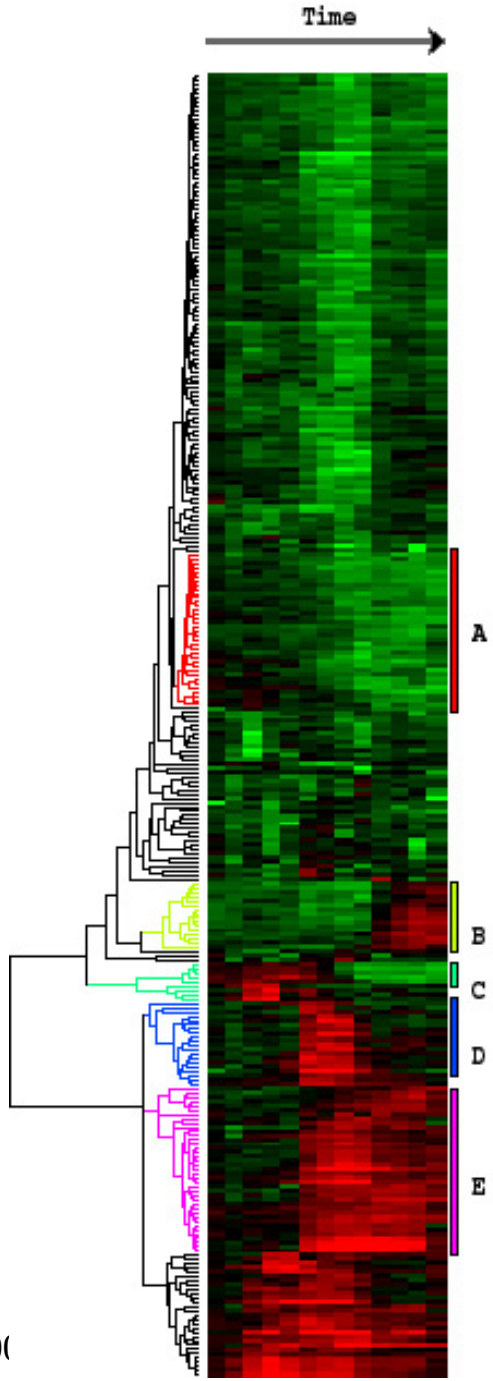


Yeast array

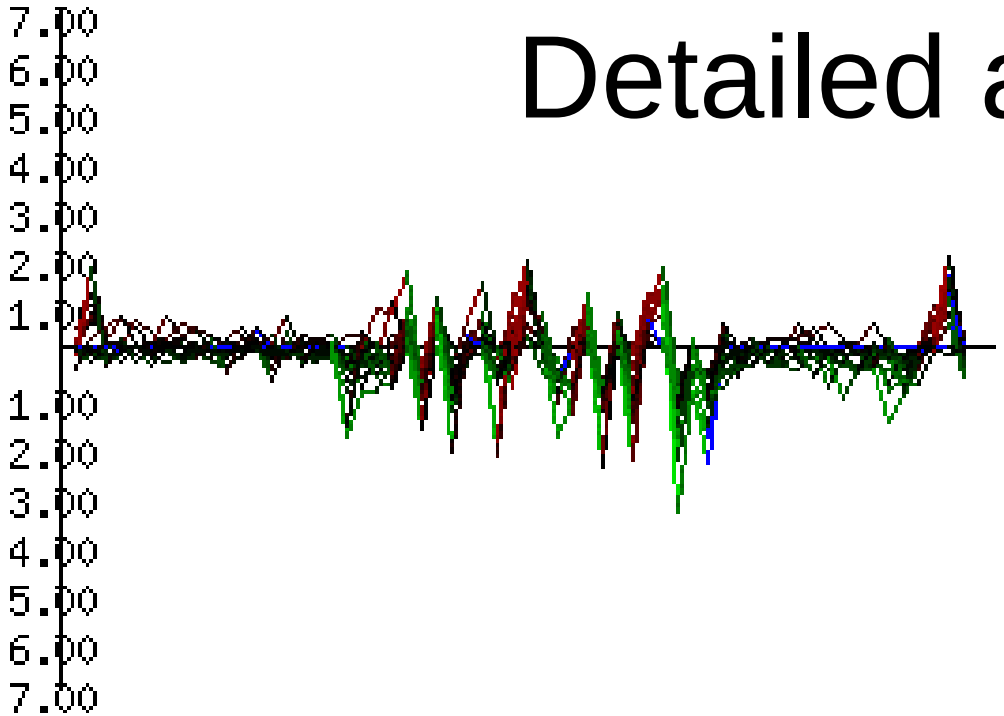


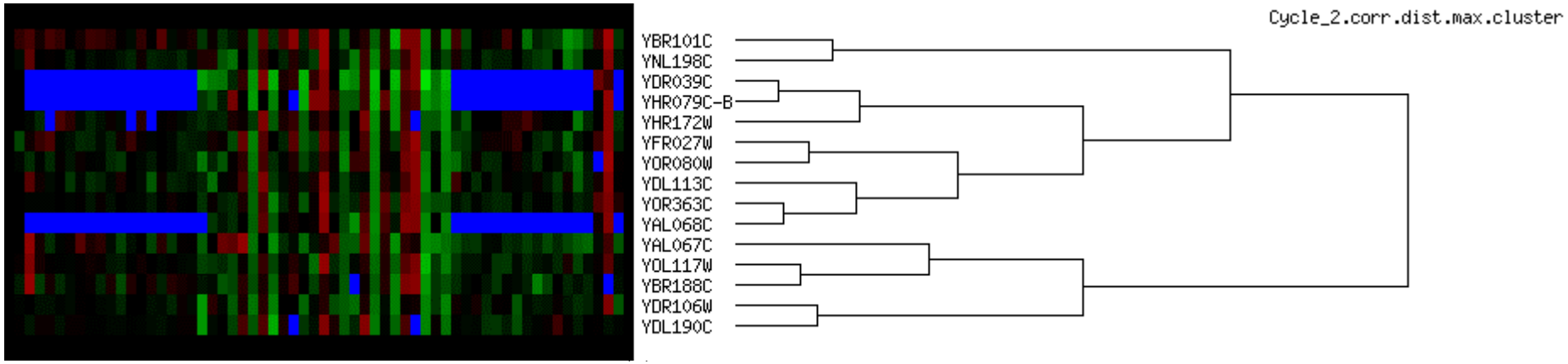

YEAST GENOME
COMPLETED
4/25/97
TOTAL TIME: 51 HRS.
SHOTS: 701,520

Analysis of array data

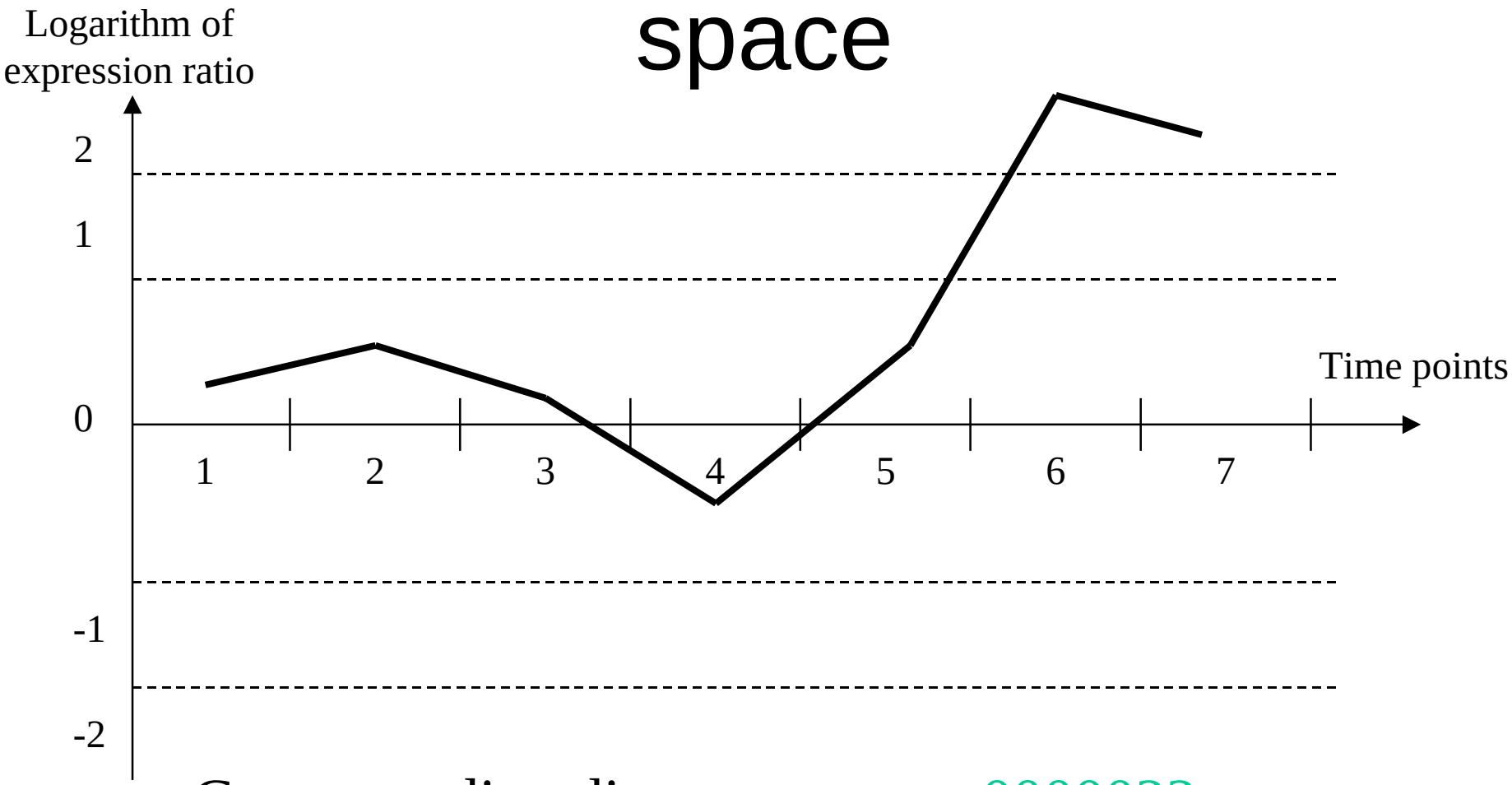


Detailed analysis...





Discretizing the gene expression measurement space



Corresponding discrete pattern: 0000022

More data related issues...

- Database design for biological resources
- Representation & visualisation of biological knowledge
- Application of data analysis methods e.g. data mining.

Data abstraction: imperative that the operations over the abstract data preserve the biological meaning of the operations on the original form of the data!

Skills and people

- Joint effort from researchers in both fields.
- Use a common language
- Learn about issues from the other side.
- Bioinformaticians specialist knowledge in maths and stats (Hidden Markov Models)
- Computer scientists: apply problem abstraction & efficient algorithm design.

Getting involved

- Work alongside with molecular biologists
- Check out what has been done before
- Validate your results
- Speed vs accuracy

But BI is great fun and a fast moving area!

Resources

www.soi.city.ac.uk/~drg/bioinformatics/resources.html

- European Bioinformatics Institute **www.ebi.ac.uk**
- National Center for Biotechnology Information **www.ncbi.nlm.nih.gov**
- Protein Data Bank **www.rcsb.org/pdb**
- Swiss-Prot Database **www.expasy.ch/sprot/sprot-top.html**
- CATH Database of Folds **www.biochem.ucl.ac.uk/bsm/cath**
- SCOP Database **scop.mrc-lmb.cam.ac.uk/scop**
- DALI **www2.ebi.ac.uk/dali**
- Structural Genomics **www.structuralgenomics.org**
- 3D Search **gene.stanford.edu/3dsearch**
- Bioinformatics course: **cmgm.stanford.edu/biochem201**
- The Bioinformatics Resource **www.hgmp.mrc.ac.uk/CCP11**
- Pattern discovery **industry.ebi.ac.uk/~brazma/patterns.html**
- Metabolic: **www.ebi.ac.uk/research/pfmp**

READING

www.soi.city.ac.uk/~drg/bioinformatics/reading.html

- T Attwood and D J Parry-Smith, An introduction to Bioinformatics, Longman 1999
- P. Baldi and S. Brunak, Bioinformatics: the machine learning approach, MIT press, 1998
- C. Branden and J. Tooze,. Introduction to Protein Structure, Second Edition, Garland Publishing, New York. http://www.proteinstructure.com/
- Creighton, T. E. (1993). Proteins: Structures and Molecular Properties. (Second Edition ed.). New York: Freeman.
- Creighton, T. E. (Ed.). (1992). Protein Folding. New York: W. H. Freeman & Co.
- Darby, N. J., & Creighton, T. E. (1993). Protein Structure. Oxford: IRL Press.

READING

- R. Durbin, S.Eddy, A. Krough and G. Mitchison, Biological Sequence Analysis, CUP 1998
- Dan Gusfield, Algorithms on strings, trees and sequences, CUP, 1997
- David B. Searls. The computational linguistics of biological sequences. In Larry Hunter, editor, Artificial Intelligence and Molecular Biology, chapter 2, pages 47-120, AAAI Press, 1993.
- David B. Searls. String Variable Grammar: A Logic Grammar Formalism for the Biological Language of DNA, Journal of Logic Programming 24:1-2, pages 73-102, 1995
- Schulz, G. E., & Schirmer, R. H. (1985). Principles of Protein Structure. New York: Springer-Verlag.
- Stryer, L. (1995). Biochemistry. (Fourth Edition ed.). New York: W. H. Freeman & Co
- Michael Waterman. Introduction to Computational Biology, Chapman & Hall, 1995