

Biological datasets management: an introduction

Luigi Palopoli
DEIS, Università della Calabria
palopoli@deis.unical.it

SEBD '05, Bressanone

1

Introduction

- ✓ Biology has been a key field for science in the 20th century
- ✓ One of the core factors pushing biological research has been and will probably be its relationships with medicine
- ✓ In this context, the recent completion of the human genome sequencing has opened new and exciting challenges
- ✓ Human genome data are just one example of the enormous amount of biological information gathered in recent times, including DNA, RNA and protein sequences in humans and several other species, structural and functional information

2

Introduction

- ✓ Most of those data, however, need still to be interpreted and their size strongly calls for tools lying in the realm of computer science: this is where Bioinformatics comes into play
- ✓ Bioinformatics is naturally an interdisciplinary field, where collaborative work amongst computer and biology experts is mandatory
- ✓ Collaboration with biologists is often not as easy, but needs to be attained for useful tools to be developed and applied to real and interesting biological problems

3

Introduction

- ✓ A definition:
"Bioinformatics is the combination of biology and information technology. It is the branch of science that deals with computer-based analysis of large biological data sets. Bioinformatics incorporates the development of databases to store and search data, and statistical tools and algorithms to analyze and determine relationships between biological data sets, such as macromolecular sequences, structures, expression profiles and biochemical pathways."
(R.M. Twyman)
- ✓ In most cases, computer based tools developed in bioinformatics require expert human intervention for the addressed problems to get solved

4

Introduction

- ✓ Useful software tools might be quite diverse, ranging from data integration middlewares to specialized dataming suites to efficient string alignment packages
- ✓ In almost all the cases, though, efficiency, flexibility and reliability are main issues: biological dataset are typically large ones and sometimes noisy
- ✓ It is also important to note that (fortunately!) most relevant biological data sets as well as several software tools for biosequences analysis are publicly available and accessible through the Web (and I'll provide several links in the sequel)

5

Introduction

- ✓ There are several facts about biology that are important to keep in mind:
 - In biology there are no rules without exceptions
 - In reasoning with biological structures, looking for generalizations maybe often misleading
 - It is often impossible to look at a biological phenomenon in isolation, for it may take place just as long as other related phenomena take place as well, which need to be taken care of too
 - To reason with incomplete information is quite the rule rather than the exception
 - In reasoning about biological structures and functions it is important to bear in mind the pervasive role of evolution

6

Introduction

- ✓ This tutorial is intended to provide a quite-in-the large overview of the field of bioinformatics
- ✓ While discussing various points I won't go in so much technical detail, except perhaps in few passages
- ✓ Also, the research area is quite articulated and large and, therefore, I won't be able to cover all possible subjects either

7

Outline

Along the presentation, I will go through the following main points:

1. Introduction
2. Some basic molecular biology
3. Biological datasets formats and repositories
4. Research goals in bioinformatics
5. Overview of problems and techniques I: Sequence data
 - a. Simple and multiple alignments
 - b. Pattern extraction from sequences
 - c. An example pattern extraction technique
6. Overview of problems and techniques II: Structural data
 - a. Determining macromolecule structure
 - b. Protein structure prediction
 - c. RNA structure prediction
7. Some further issues
8. Conclusion

8

Some basic molecular biology

- ✓ This part of the tutorial is intended to provide some few concepts about basic mechanisms underlying cell functioning
- ✓ The Web is a enormous source of information about the issues we are going to briefly touch next
- ✓ Also, several text books are available to gain familiarity with those subjects

9

Some basic molecular biology

Cells

- ✓ Living cells consist of macromolecular components (DNA, proteins)
- ✓ Such molecules interact with each other, by:
 - Joining, thus forming new molecules
 - Splitting, thus disassembling into fragments
- ✓ Most often, interaction may take place only as a result of the existence of some specifical shape and location of chemical constituents of involved molecules

10

Some basic molecular biology

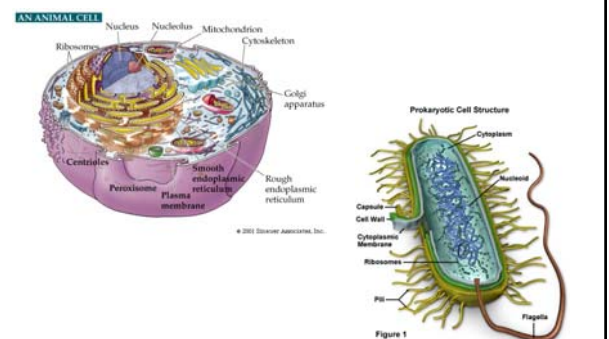
Cells

- ✓ Such interactions determine and regulate the lifecycle of cells
- ✓ For instance, prokaryotic cells grow because nutrients coming from the outside pass through cell's membrane and participate in interactions with internal components; components are also expelled by the cell
- ✓ Proteins have multiple roles and are the core structures influencing cell's life (enzymatic processes, regulatory function,...)

11

Some basic molecular biology

Cells



12

Some basic molecular biology

Cells

- ✓ Besides growing, cells also reproduce themselves by cell division
- ✓ DNA duplication capability has a central role in allowing cell division (as well as protein synthesis)
- ✓ There are other important characteristics of the living cells, such as motility and death mechanisms

13

Some basic molecular biology

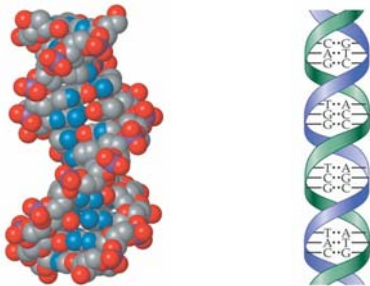
DNA

- ✓ DNA is a complex helix-shaped structure consisting of two strands of *Nucleotides*
- ✓ There are four nucleotides found in DNA:
 - A – Adenine
 - C – Cytosine
 - G – Guanine
 - T – Thymine
- ✓ DNA strands are *complementary*: for each strand S, each occurrence of A in S is paired with a T in the other strand, and C with G; so, it suffices to specify the sequence of one strand to get both strands' sequences

14

Some basic molecular biology

DNA



15

Some basic molecular biology

Amino Acids

- ✓ *Proteins* are the core structures determining cell lifecycle; they are made up of 20 *aminoacids* (few exceptions exist) or *residues*

1. A	Ala	Alanine
2. C	Cys	Cysteine
3. D	Asp	Aspartic Acid
4. E	Glu	Glutamic Acid
5. F	Phe	Phenylalanine
6. G	Gly	Glycine
7. H	His	Histidine
8. I	Ile	Isoleucine
9. K	Lys	Lysine
10. L	Leu	Leucine

16

Some basic molecular biology

Amino Acids

✓ Residues ... (continued)

11. M	Met	Methionine
12. N	Asn	Asparagine
13. P	Pro	Proline
14. Q	Gln	Glutamine
15. R	Arg	Arginine
16. S	Ser	Serine
17. T	Thr	Threonine
18. V	Val	Valine
19. W	Trp	Tryptophan
20. Y	Tyr	Tyrosine

17

Some basic molecular biology

DNA and proteins

- ✓ There exists a wide variety of proteins in living organisms, but they always have those aminoacids as their building constituents
- ✓ *DNA* contains all the information needed for protein synthesis (*DNA coding regions*)
- ✓ Each residue in fact translates from one or more triplets of nucleotides making up DNA; for instance:

GAA → E (Glu – Glutamic Acid)
AAC → N (Asn – Asparagine)

18

Some basic molecular biology

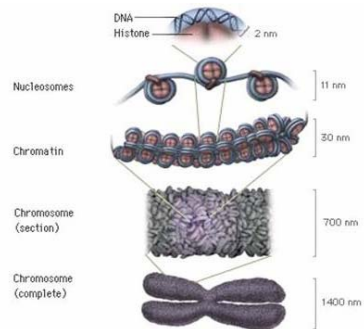
DNA and proteins

- ✓ DNA coding regions (lying in either strands) are called *genes*; the set of genes as well as intergenic regions in a given organism is called its *genome*
- ✓ The role of the genetic material outside genes is largely unknown (there is plenty of it, especially in complex beings – coding regions in humans represent less than 3% of the human genetic material – overall about 3 billions base pairs)
- ✓ However, it is known that some regions do play regulatory functions

19

Some basic molecular biology

DNA and proteins



20

Some basic molecular biology

Gene expression

- ✓ *Expressing a gene* means that that gene is activated to produce RNA (the amount of such RNA is proportional to the quantity of the corresponding protein)
- ✓ This can be used, for instance, in order to single out abnormal genes responsible, e.g., for anomalous overproduction of a given protein within a specific disease context

25

Some basic molecular biology

Protein structures

- ✓ Once produced, a protein *folds* into a 3D structure, which is important in determining its function, that might be rather diverse (for instance, may either cause or stop the production of other proteins)
- ✓ For most proteins (globular proteins) the folding is dominated by the hydrophobic effect determined by the characteristics of some amino acids

26

Some basic molecular biology

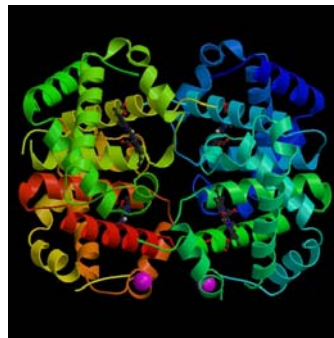
Protein structures

- ✓ Globular proteins contain elements of regular secondary structures, like α -helices and β -sheets (or strands); residues lying within such regular substructures are often marked H (helix) and E or B (sheet)
- ✓ Membrane proteins (that exist within lipid membranes) are also characterized by the same secondary structure topologies, but their folding obey different structural principles than globular proteins

27

Some basic molecular biology

Protein structures



An image of the folding of 4HHB (Human Hemoglobin)

28

Some basic molecular biology

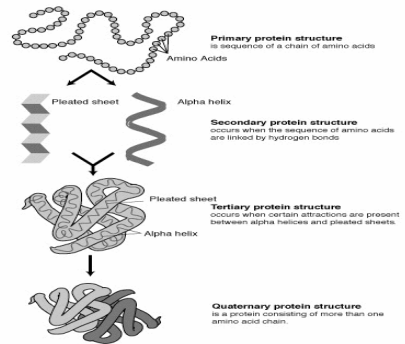
Protein structures

- ✓ The tertiary structure of a protein is the full 3D atomic structure of its components. This will comprise residues belonging to regular secondary structures plus those organized in irregular loops, mostly lying on the protein surface. Loops residues are often marked by C
- ✓ Several tertiary structures, packed together, may finally form the biological functional quaternary structure

29

Some basic molecular biology

Protein structures



30

Some basic molecular biology

Protein domains

- ✓ Functionally, a protein can be decomposed into *domains*: each domain is a "simple" functional unit which corresponds, from the structural viewpoint, to a portion of the protein sequence able to fold independently of the rest
- ✓ Each domain may occur in various, evolution-unrelated proteins
- ✓ For instance, the *Homeo* domain (serves the purpose of binding DNA) is found in human PAX4 and IHX2 and in mouse Cux-2

31

Some basic molecular biology

Protein domains

- ✓ All the structures I have mentioned (as well as whole organisms) are the results of the work done by evolution through times
- ✓ In particular, at the sequence level evolution accepts changes to a basis or residue if this have a neutral or advantageous effect
- ✓ For instance, in proteins this means that the change has not to compromise the protein structural stability nor its function.

32

Some basic molecular biology

Protein domains

- ✓ When comparing related sequences (homologous) it is observed that key structural and functional components are usually conserved
- ✓ The conservation of structure and function can go far beyond sequence similarity recognition
- ✓ There are cases also where structure is conserved, whereas function is not (e.g., with gene duplication); in this case a stability of structural residues would be observed along with changings in the functional ones

33

Some basic molecular biology

Protein domains

- ✓ Summarizing, amongst other relevant things, the following are important ones to recall when talking about biosequences analysis:
 - The central dogma
 - The role of evolution
 - The shape assumed by macromolecules in the 3D space

34

Outline

Along the presentation, I will go through the following main points:

1. Introduction
2. Some basic molecular biology
3. Biological datasets formats and repositories
4. Research goals in bioinformatics
5. Overview of problems and techniques I: Sequence data
 - a. Simple and multiple alignments
 - b. Pattern extraction from sequences
 - c. An example pattern extraction technique
6. Overview of problems and techniques II: Structural data
 - a. Determining macromolecule structure
 - b. Protein structure prediction
 - c. RNA structure prediction
7. Some further issues
8. Conclusion

35

Biological datasets formats and repositories

- ✓ Biologically relevant macromolecules are naturally modeled as 3D structures
- ✓ However, the simplification is often to be introduced to look at them as strings of symbols (nucleotides or aminoacids) from where to eventually reconstruct 3D structures
- ✓ Most of the available data have been obtained by sequencing techniques and encoded in strings
- ✓ DNA has a known 3D structure, whereas the problem of shape determination from sequences is crucial for RNA and proteins

36

Biological datasets formats and repositories

- ✓ Available information about 3D structures has a much smaller size than that describing sequences
- ✓ A recent survey tells us that NCBI servers (www.ncbi.nlm.nih.gov) report 26 Giga base pairs representing various genomes (about 3G bp for the human genome) with the largest gene comprising 20M bp and the largest protein consisting of about 34K aminoacids
- ✓ The protein databank PDB (www.rcsb.org/pdb/) reports about 30K proteins listed along with their 3D structures

37

Biological datasets formats and repositories

- ✓ There are conventions for representing nucleic acids and protein sequences, among which the following are widely used:
 - NBRF/PIR
 - FASTA
 - GDE
- ✓ Data are usually stored in text files where spaces and CR are *ignored*

38

Biological datasets formats and repositories

- ✓ The formats listed above allow to represent, along with the sequences themselves, also unique identification codes and comments typically comprising the sequence name, the species from which that certain sequence was derived and accession numbers to genetic data banks

39

Biological datasets formats and repositories

- ✓ The NBRF/PIR format:

```
>P1; 5H1B_CAVPO
Guinea pig serotonin receptor accession: 008892
MGNPEASCTP PAVLGSQTGL PHANVSAPPN ...
ATTLNSAFVI ATVYRTRKLH TPANY LIASL ...
.....
*
sequence
```

- ✓ Extensions “.pir” or “.seq”

40

Biological datasets formats and repositories

- ✓ The FASTA format:

```
>5H1B_CAVPO 008892 | guinea pig serotonin receptor
MGNPEASCTP PAVLGSQTGL PHANVSAPPN ...
ATTLSNAFVI ATVYRTRKLH TPANY LIASL ...
.....
```

- ✓ Extension ".fasta"

41

Biological datasets formats and repositories

- ✓ The GDE format:

```
%5H1B_CAVPO 008892 | guinea pig serotonin receptor
MGNPEASCTP PAVLGSQTGL PHANVSAPPN ...
ATTLSNAFVI ATVYRTRKLH TPANY LIASL ...
.....
```

- ✓ Extension ".gde"

42

Biological datasets formats and repositories

- ✓ A major product of bioinformatics tools are sequence alignments (which I will talk about shortly)
- ✓ NBRF/PIR, FASTA and GED can be used to represent aligned sequences
- ✓ However, more specialized formats exist, like ALN produced by CLUSTALW/X or MSF

43

Biological datasets formats and repositories

```
MSF: 435      Type: P      Check: 2299      .....
Name: 5H1A_MOUSE oo Len: 435 Check: 7521 Weight: 0.166
Name: 5H1A_RAT oo Len: 435 Check: 7521 Weight: 0.166
Name: 5H1A_HUMAN oo Len: 435 Check: 7521 Weight: 0.166
Name: 5H1A_CAVPO oo Len: 435 Check: 7521 Weight: 0.166
Name: 5H1A_CRIGR oo Len: 435 Check: 7521 Weight: 0.166
```

```
5H1A_MOUSE      MD ..... MF SLGQGNNTTT .....
5H1A_RAT        MD ..... VF SFGQGNNTTA .....
5H1A_HUMAN      MD ..... VL SPGQGNNTTS .....
5H1A_CAVPO      MGNPEASCTP .....
.....           .....

```

← Aligned sequences with gaps

44

Biological datasets formats and repositories

- ✓ A further important data format category groups those storing structural information
- ✓ The most notable example is that of the Protein Data Bank (www.rcsb.org/pdb)
- ✓ The PDB delivers text files where the structural features (as well as other informations) about proteins and other compounds are described

45

Biological datasets formats and repositories

- ✓ In a PDB file, ATOM lines tell where molecules are located in terms of a 3D space
- ✓ Aminoacids are represented using their three-letters codes
- ✓ An example query result from PDB is reported next

46

Biological datasets formats and repositories PDB search for 4HHB (1)

```
HEADER OXYGEN TRANSPORT 07-MAR-84 4HHB
COMPND HEMOGLOBIN (DEOXY)
SOURCE HUMAN (HOMO SAPIENS)
AUTHOR G.FERMI,M.F.PERUTZ
REVDAT 2 15-OCT-89 4HHBA 3 MTRIX
REVDAT 1 17-JUL-84 4HHB 0
SPRSDE 17-JUL-84 4HHB 1HHB
JRNL AUTH G.FERMI,M.F.PERUTZ,B.SHAANAN,R.FOURME
JRNL TITL THE CRYSTAL STRUCTURE OF HUMAN.....
JRNL REF J.MOL.BIOL. V. 175 159 1984
JRNL REFN ASTM JMOBAKUK ISSN 0022-2836 070
REMARK 1 REFERENCE 1
REMARK 1 AUTH M.F.PERUTZ,S.S.HASNAIN,P.J.DUKE.....
REMARK 1 TITL STEREOCHEMISTRY OF IRON .....
REMARK 1 REF NATURE V. 295 535 1982
REMARK 1 REFN ASTM NATUAS UK ISSN 0028-0836 006
REMARK 1 REFERENCE 2
```

47

Biological datasets formats and repositories PDB search for 4HHB (2)

```
SEQRES 1A 141 VAL LEU SER PRO ALA ASP LYS THR ASN VAL
LYS ALA ALA
SEQRES 2A 141 TRP GLY LYS VAL GLY ALA HIS ALA GLY GLU
TYR GLY ALA
.....
FTNOTE 1
FTNOTE 1 PROBABLY PHOSPHATE GROUP.
.....
FORMUL 5 HEM 4(C34 H32 N4 O4 FE1 ++)
```

48

Biological datasets formats and repositories

PDB search for 4HHB (3)

```
.....
.....
ATOM   1  N  VAL A  1  6.204 16.869 4.854 7.00 49.05
ATOM   2  .....
.....
TER      1070  ARG  A 141
HETATM  1071  FE   HEM A 1  8.116 7.403 -15.045 24.00 18.07
.....
CONNECT  650 648 649 1071
.....
MASTER  94  2  6  32  0  0  0  9 4779  4 180 46
END
```

49

Biological datasets formats and repositories

- ✓ Primary data bases are those containing raw sequence data, such as GenBank (www.ncbi.nlm.nih.gov/GenBank/index.html) and the NSD (EMBL – www.ebi.ac.uk/embl/)
- ✓ An example query result taken from GenBank is shown in the following slides

50

Biological datasets formats and repositories

Genbank search for BTEB (1)

```
LOCUS      NM_001206 5208 bp mRNA linear PRI 14-MAY-2005
DEFINITION Homo sapiens Kruppel-like factor 9 (KLF9), mRNA.
ACCESSION NM_001206VERSION NM_001206.2 GI:59853224
KEYWORDS
SOURCE     Homo sapiens (human)
ORGANISM   Homo sapiens
           Eukaryota; .....
REFERENCE  1 (bases 1 to 5208)
AUTHORS    Zhang,X.L., .....
TITLE      Selective interactions of Kruppel-like factor 9/basic .....
JOURNAL    J. Biol. Chem. 278 (24), 21474-21482 (2003)
PUBMED     12672823
REFERENCE  2 (bases 1 to 5208).....
```

51

Biological datasets formats and repositories

Genbank search for BTEB (2)

```
FEATURES   Location/Qualifiers
source     1..5208
           /organism="Homo sapiens"
           /mol_type="mRNA"
           /db_xref="taxon:9606"
           /chromosome="9"
           /map="9q13" gene 1..5208
           /gene="KLF9"
           /note="synonyms: BTEB, BTEB1"
           /db_xref="GeneID:687"
           .....
           /gene="KLF9"
           /note="basic transcription element binding protein 1;
           go_component: nucleus [goid 0005634] [evidence IEA];
           go_function: zinc ion binding [goid 0008270][evidence IEA];
           .....
           go_process: transcription [goid 0006350] [evidence IEA];
           .....
```

52

Biological datasets formats and repositories

Genbank search for BTEB (3)

```
/codon_start=1
/product="Kruppel-like factor 9"
/protein_id="NP_001197.1"
/db_xref="GI:4557375"
/translation="MSAAAYMDFVAAQCLVSI SNRAAVPEHG....."
polyA_signal 5180..5185
/gene="KLF9"
polyA_signal 5189..5194
/gene="KLF9"
polyA_site 5208
/gene="KLF9"
ORIGIN
1 cttactcatt tgtgtttatt ctggactta tctgacata atggggtt.....
```

53

Biological datasets formats and repositories

- ✓ Other database are called *secondary* because they maintain data elaborated starting from raw data
- ✓ Two notable secondary databases are SWISS-PROT (www.expasy.org/sprot) and the related UniProtKB/TrEMBL resource
- ✓ SWISS-PROT is a collection of confirmed protein sequences with annotations relating to structure, function and protein family; TrEMBL contains translations of EMBL nucleotide sequences entries not yet in SWISS-PROT

54

Biological datasets formats and repositories

- ✓ Those databases can be searched by sequence similarity
- ✓ Also, text-based searches are supported on annotations
- ✓ An example result query taken from SWISS-PROT is reported in the following slides

55

Biological datasets formats and repositories

Swiss-Prot search for BTEB (1)

```
ID BTEB1_HUMAN STANDARD; PRT; 244 AA.
AC Q13886; Q16196;
DT 15-DEC-1998 (Rel. 37, Created)
.....
DE Transcription factor BTEB1
GN Name=KLF9; Synonyms=BTEB, BTEB1;
OS Homo sapiens (Human).
OC Eukaryota; .....
OX NCBI_TaxID=9606;
RN [1]
RP NUCLEOTIDE SEQUENCE.
RX MEDLINE=94120483; PubMed=8291025[NCBI, ExPASy, EBI...];
RA Ohe N., Yamasaki Y., .....;
RT "Chromosomal localization and cDNA sequence of human BTEB
....."
RL Somat. Cell Mol. Genet. 19:499-503(1993).
RN [2]
.....
```

56

Biological datasets formats and repositories

Swiss-Prot search for BTEB (2)

DR EMBL; D31716; BAA06524.1; -; mRNA. [EMBL / GenBank / DDBJ] [CoDingSequence]
DR PIR; I59602; I59602.
DR HSSP; P08047; 1SP2. [HSSP ENTRY / SWISS-3DIMAGE / PDB]
DR TRANSFAC; T02212; -.
DR Ensembl; ENSG00000119138; Homo_sapiens
DR Ensembl; HGNC:1123: KLF9.
DR CleanEx; HGNC:1123: KLF9.
DR MIM; 602902; -, [NCBI / EBI]
DR GeneCards; KLF9.
DR GeneLynx; KLF9.
DR GenAtlas; KLF9.
DR SOURCE; KLF9.
DR GO; GO:0003700; F:transcription factor activity; TAS.
DR InterPro; IPR007087; Znf_C2H2.
DR InterPro; Graphical view of domain structure.
.....

57

Biological datasets formats and repositories

Swiss-Prot search for BTEB (3)

KW DNA-binding; Metal-binding; Nuclear protein; Repeat; Transcription;
KW Transcription regulation; Zinc; Zinc-finger.
FT ZN_FING 143 167 C2H2-type 1.
FT ZN_FING 173 197 C2H2-type 2.
FT ZN_FING 203 225 C2H2-type 3.
FT COMPBIAS 84 116 Asp/Glu-rich (acidic).
SQ SEQUENCE 244 AA; 27235 MW; 2D1B5A5BB9D42221 CRC64;

MSAAAYMDFV AAQCLVSISSN RAAVPEHGVA PDAERLRLPE
REVTKHEGDP GDTWKDYCTL VTIAKSLLDL NKYRPIQTPS
VCSDSLESPD EDMGSDSDVT TESGSSPSHS PEERQDPGSA
.....

58

Biological datasets formats and repositories

- ✓ Besides general databases as those cited before, many organism-specific databases have been organized and are maintained; their number grows as more genome projects are started
- ✓ A much useful Web gateway to genome databases is GOLD (www.genomesonline.org): it lists 266 complete genomes, 730 prokaryotic ongoing genome sequencings and 496 eukaryotic ongoing genome sequencings
- ✓ Such databases contain not only sequence data, but also information on gene expression, mutant phenotypes and relevant scientific literature

59

Biological datasets formats and repositories

- ✓ Available sequenced genomes comprise, for instance, those of the following organisms:
 - Saccharomyces cerevisiae
 - Drosophila melanogaster
 - Mouse
 - Human
- ✓ Also, databases are maintained that store information about specific data
- ✓ Examples are OMIM (On Line Mendelian Inheritance in Man – www.ncbi.nlm.nih.gov/omim) and InBase, dealing with small peptides (www.neb.com/neb/inteind.html)

60

Biological datasets formats and repositories

- ✓ Finally, other databases serve specific research aspects, such as higher-level functions:
 - PIR (Protein Information Resource at pir.georgetown.edu)
 - PATHWAY and LIGAND linked from KEGG (Kyoto Encyclopedia of Genes and Genomes at www.genome.ad.jp/kegg)
 - DIP (Database of Interacting Proteins at dip.doe-mbi.ucla.edu) and the general gateway PID (Protein Interaction Databases at www.hgmp.mrc.ac.uk/GenomeWeb/prot-interaction.html)

61

Biological datasets formats and repositories

- ✓ Other databases (continued):
 - PROSITE that contains patterns associated to protein families (patterns obtained via multiple alignments); found at www.expasy.org/prosite
 - PRINTS (at umber.sbs.man.ac.uk/dbbrowser/PRINTS) representing protein families as multiply aligned ungapped segments derived from most highly conserved regions in the family
 - CATH (at cathwww.biochem.ucl.ac.uk/latest/) and SCOP (at scop.mrc-lmb.cam.ac.uk/scop/) storing protein structure classifications

62

Biological datasets formats and repositories

- ✓ Before closing this section of the tutorial it is worth pointing out that several biological database services allows for the query results to be delivered in XML format besides native formats as the ones we have shown before
- ✓ This is true, for instance, for GenBank and PDB (a PDB XML example result is shown next)

63

Biological datasets formats and repositories

PDB search for 4HHB (1) – XML output (partial)

```
<?xml version="1.0" encoding="UTF-8" ?>
<PDBx:datablock datablockName="4HHB"
  xmlns:PDBx="http://deposit.pdb.org/pdbML/pdbx.xsd"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://deposit.pdb.org/pdbML/pdbx.xsd">
  <PDBx:atom_sites_footnoteCategory>
    <PDBx:atom_sites_footnote id="1">
      <PDBx:text>
        PROBABLY PHOSPHATE GROUP.
      </PDBx:text>
    </PDBx:atom_sites_footnote>
  </PDBx:atom_sites_footnoteCategory>
  <PDBx:audit_authorCategory>
    <PDBx:audit_author name="Fermi, G."></PDBx:audit_author>
    <PDBx:audit_author name="Perutz,MF."></PDBx:audit_author>
  </PDBx:audit_authorCategory>
```

64

Biological datasets formats and repositories

PDB search for 4HHB (2) – XML output (partial)

```
<PDBx:cellCategory>
  <PDBx:cell_entry_id="4HHB">
    <PDBx:length_a>63.150</PDBx:length_a>
    <PDBx:length_b>83.590</PDBx:length_b>
    <PDBx:length_c>53.800</PDBx:length_c>
    <PDBx:angle_alpha>90.00</PDBx:angle_alpha>
    <PDBx:angle_beta>99.34</PDBx:angle_beta>
    <PDBx:angle_gamma>90.00</PDBx:angle_gamma>
    <PDBx:Z_PDB>4</PDBx:Z_PDB>
  </PDBx:cell>
</PDBx:cellCategory>
```

65

Biological datasets formats and repositories

PDB search for 4HHB (3) – XML output (partial)

```
<PDBx:database_PDB_matrixCategory>
  <PDBx:database_PDB_matrix entry_id="4HHB">
    <PDBx:origx11>.963457</PDBx:origx11>
    <PDBx:origx12>.136613</PDBx:origx12>
    <PDBx:origx13>.230424</PDBx:origx13>
    <PDBx:origx21>-.158977</PDBx:origx21>
    <PDBx:origx22>.983924</PDBx:origx22>
    <PDBx:origx23>.081383</PDBx:origx23>
    <PDBx:origx31>-.215598</PDBx:origx31>
    <PDBx:origx32>-.115048</PDBx:origx32>
    <PDBx:origx33>.969683</PDBx:origx33>
    <PDBx:origx_vector1>16.61000</PDBx:origx_vector1>
    <PDBx:origx_vector2>13.72000</PDBx:origx_vector2>
    <PDBx:origx_vector3>37.65000</PDBx:origx_vector3>
  </PDBx:database_PDB_matrix>
</PDBx:database_PDB_matrixCategory>
```

66

Biological datasets formats and repositories

PDB search for 4HHB (4) – XML output (partial)

```
<PDBx:struct_refCategory>
  <PDBx:struct_ref id="1">
    <PDBx:db_name>SWS</PDBx:db_name>
    <PDBx:db_code>HBA_HUMAN</PDBx:db_code>
    <PDBx:entity_id>1</PDBx:entity_id>
    <PDBx:pdbx_db_accession>
      P01922
    </PDBx:pdbx_db_accession>
    <PDBx:pdbx_align_begin>1</PDBx:pdbx_align_begin>
    <PDBx:pdbx_seq_one_letter_code>
      VLSPADKTNVKAAWGKVGAGHAGEYGAELERMFLSFP
      TTKTYFPFDLSHGSAQVKGHGKKVADA.....
    </PDBx:pdbx_seq_one_letter_code>
  </PDBx:struct_ref>
  .....
</PDBx:struct_refCategory>
  .....
</PDBx:datablock>
```

67

Outline

Along the presentation, I will go through the following main points:

1. Introduction
2. Some basic molecular biology
3. Biological datasets formats and repositories
4. Research goals in bioinformatics
5. Overview of problems and techniques I: Sequence data
 - a. Simple and multiple alignments
 - b. Pattern extraction from sequences
 - c. An example pattern extraction technique
6. Overview of problems and techniques II: Structural data
 - a. Determining macromolecule structure
 - b. Protein structure prediction
 - c. RNA structure prediction
7. Some further issues
8. Conclusion

68

Research goals in bioinformatics

- ✓ Generally speaking, the aim of bioinformatics is to help biologists in gathering and processing biological data and to aid in studying protein structures and interactions in order to allow optimal drug design
- ✓ Tasks typically comprise:
 - Determining proteins' 3D structure and functions starting from their aminoacid sequences
 - Determining genes and proteins associated to a given genome
 - Determining drug binding sites in proteins

69

Research goals in bioinformatics

- ✓ A main methodological tool in most of those problems is *searching for homologues*: homology between two sequences or two structures suggests for the existence of common ancestors and, this, most often, indicates similar functions and mechanisms in turn
- ✓ Even if often so, it must be noted that not always similarity in sequences translates into similarity in structures
- ✓ In any case, similarity search is central to most bioinformatic application domains

70

Research goals in bioinformatics

- ✓ Within the context outlined before, several issues are good subjects for research in computer science in general and databases in particular, with good potentialities for application to biology
- ✓ I will list several of them next

71

Research goals in bioinformatics

- ✓ *Algorithms and indexing methods for comparing sequences*, with focus on:
 - *efficiency*, since huge amounts of significantly long sequence are there to be analyzed
 - *flexibility*, since evolution makes it allowable for ins, replacement and del to occur in compared sequences and, moreover, it is often the case that one does not exactly know what to look for

72

Research goals in bioinformatics

- ✓ *Techniques for classifying sequences and structures*, useful in order to construct phylogenies and distant evolutionary relationships: datamining techniques such as specialized clustering, classification and outlier detection methods are supposedly very useful in this realm
- ✓ *Predicting 3d structures from sequences*, is a fundamental problem to be dealt with and a computational difficult ones; homology-based strategies can be adopted as well as ab-initio methods; also, approaches allowing for diverse prediction strategies to be integrated can be valuable

73

Research goals in bioinformatics

- ✓ *Pattern extraction techniques*, there are parts of DNA and aminoacid sequences that need to be recognized and extracted (note: it is not pattern matching!); examples are:
 - determining genes and other active loci in nucleic sequences;
 - identifying unusual or over-represented subsequences;
 - recognizing structure-constrained repetitive patterns (for they may mark conserved or anomalous sequence regions);machine learning, statistical, automata-theoretic, data mining and indexing techniques seems to be usefully applied in this context

74

Research goals in bioinformatics

- ✓ *Inferring cell regulation*, is a formidable problem involving modeling and simulating cell regulation mechanisms; machine learning and other inductive techniques applied over experimental data (notably, microarray samples) seem to be promising
- ✓ *Determining protein functions and pathways*, again, a difficult problem and a relevant one for which not much data are readily available; the problem is to interpret experts' annotations found in protein databases about functions and interactions and produce integrated databases to be queried for the existence of specific reactions or chains of reactions involving proteins; database integration, wrapping and text mining techniques are, probably, ways to go

75

Research goals in bioinformatics

- ✓ *Database and information services integration*, a more classical research subject for database people but, again, one with a potential relevant impact on the molecular biology community; wrapping and mediation techniques, incomplete data, Web services, workflow management techniques might be conveniently applied
- ✓ Last, but not the least, all those *specific problems* that arise while collaborating with biologists that often require Sw tools to be improved or utilized in novel ways; an example is applying data mining techniques to single out protein markers of a given disease starting from profiles obtained from mass spectrometry data obtained from healthy and sick populations

76

Research goals in bioinformatics

- ✓ To close with this subject, here is a summary of CS methods and techniques relevant to bioinformatics:
 - ✓ String algorithms, grammars and automata
 - ✓ Indexing methods and query optimization
 - ✓ Integration techniques
 - ✓ Optimization techniques
 - ✓ Dynamic programming and heuristics
 - ✓ Data mining and machine learning techniques
 - ✓ Probability and statistic-based methods
 - ✓ Computational geometry methods
 - ✓ Text mining and NLP methods
 - ✓ GUI design techniques

77

Outline

Along the presentation, I will go through the following main points:

1. Introduction
2. Some basic molecular biology
3. Biological datasets formats and repositories
4. Research goals in bioinformatics
5. Overview of problems and techniques I: Sequence data
 - a. Simple and multiple alignments
 - b. Pattern extraction from sequences
 - c. An example pattern extraction technique
6. Overview of problems and techniques II: Structural data
 - a. Determining macromolecule structure
 - b. Protein structure prediction
 - c. RNA structure prediction
7. Some further issues
8. Conclusion

78

Problems and techniques I: Sequence data Simple alignments

- ✓ Most often biological databases have to be searched for sequences similar to a given one in order to predict the structure or the functions of the query sequence; the rationale here is that similar sequences should share a common ancestor and, as such, common structures and/or functions
- ✓ Any pair of DNA sequences will show some degree of similarity; sequence alignment is needed in order to distinguish true biological relationships from chance similarities
- ✓ Alignment algorithms use scores and gap penalties to quantify similarity

79

Problems and techniques I: Sequence data Simple alignments

Example:

A	A	T	T	G	A	T	T	G	C	G	C	A	T	T	T	A	A	G	G	G	
A	A	C	T	G	A	-	-	-	C	G	C	A	T	C	T	T	A	A	G	G	G

- ✓ Gaps are introduced in order to produce a better alignment.
- ✓ Alignments can be interpreted in evolutionary terms:
 - When letters match it means that those were part of an ancestral subsequence that remained unchanged
 - When non identical letters are aligned this means that a mutation has occurred
 - Gaps also correspond to mutations realized as insertions or deletions

80

Problems and techniques I: Sequence data

Simple alignments

- ✓ Alignment can be obtained by applying dynamic programming techniques
- ✓ Among these, the Needleman-Wunsch algorithm can be used to produce global alignments, whereas the Smith-Waterman one produces local similarity
- ✓ Gap-penalties are introduced in order to (negatively) weigh the occurrence of gaps; matching letters improve the overall score assigned to the alignment

81

Problems and techniques I: Sequence data

Simple alignments

- ✓ Several forms of gap penalties can be used: constant, length-proportional, affine (this last uses both a constant and a length-proportional contribution)
- ✓ Specific alignment scoring also depends on the sequence type: with DNA makes sense to look simply at matching symbols and gaps; with proteins not any substitution has the same biological sense, since some aminoacids are substituted by some others better than the rest of them, and this must be taken into account while looking for best alignments

82

Problems and techniques I: Sequence data

Simple alignments

- ✓ Two sequences are *homologous* only if they descend from a common ancestor
- ✓ Homologous sequences most often show similar functions: if a function is essential for an organism to live, this function will be selected by evolution and transmitted to descent

83

Problems and techniques I: Sequence data

Simple alignments

- ✓ In protein sequences, aminoacid substitutions survive evolution only if they have no deleterious effects on protein structural stability and function
- ✓ Therefore, changes occurring in protein sequences usually involve substitutions between aminoacids with similar properties:
 - ✓ Group 1 - Hydrophobic: {A, G, P, I, L, V, C, M, W, F}
 - ✓ Group 2 - Aromatic: {W, F, H, Y}
 - ✓ Group 3 - Polar: {Y, S, T, N, Q}
 - ✓ Group 4 - Basic: {H, K, R}
 - ✓ Group 5 - Acidic: {D, E}

84

Problems and techniques I: Sequence data

Simple alignments

- ✓ Substitution matrices are used to correctly weigh substitutions between aminoacids; two widely used sets of such matrices are PAMs and BLOSUMs
- ✓ A small fragment of PAM250 is shown below (identical aminoacid matching are weighed on the basis of the aminoacid occurrence frequency):

C	12		
S	0	2	
T	-2	1	3
.....	C	S	T

85

Problems and techniques I: Sequence data

Simple alignments

- ✓ Using alignment algorithms, sequence databases can therefore be searched looking for sequences which are the most similar to a given (query) sequence
- ✓ Resulting sequences are ranked on a similarity basis with respect to the query sequence
- ✓ With huge databases, even DP algorithms might turn out to be too slow
- ✓ Usually, heuristic methods are adopted, which do not guarantee optimality, but are much faster than DP algorithms

86

Problems and techniques I: Sequence data

Simple alignments

- ✓ Probably the most used is BLAST
- ✓ BLAST is faster than DP algorithms (up to 50 times faster) and in fact produces almost optimal results
- ✓ It is based on the idea of starting with short subsequences of identical letters which are then extended to a full alignment

87

Problems and techniques I: Sequence data

Simple alignments

- ✓ For simplicity, I will briefly discuss the basic BLAST algorithm, without considering gap management; variants are implemented that indeed manage gaps
- ✓ Thus, the objective here is that of identifying pairs of subsequences matching with a similarity score above a certain threshold

88

Problems and techniques I: Sequence data

Simple alignments

- ✓ Define in this context a *segment pair* to be a pair of subsequences with the same length and extracted from the two sequences under examination
- ✓ Define a *maximal local score* to be a score that is achieved on a segment pair and cannot be improved considering neither super- nor sub- segments of the given ones

89

Problems and techniques I: Sequence data

Simple alignments

The basic algorithm

1. Consider segment pairs of length w scoring at least T points, with w and T are suitable given parameter values
2. List those segments and find them within the given sequences; these occurrences form the *seeds*
3. Extend all the seeds until maximal local scores are attained

90

Problems and techniques I: Sequence data

Simple alignments

NCBI **protein-protein BLAST**

Search: adEaEgtagEgnggd

Get subsequence From: To:

Choose database: **pdb**

Do CC Search

How: **BLAST!** **BLAST!** **BLAST!**

Options for advanced blasting

Limit to species: **All organisms** or select from:

Composition based statistics

91

Problems and techniques I: Sequence data

Simple alignments

- ✓ But when an identified sequence similarity is indeed significant?
- ✓ To evaluate significance, two measures are usually adopted (and actually computed by most existing Sw packages): the *p value* and the *E value*

92

Problems and techniques I: Sequence data

Simple alignments

- ✓ The *p value* of an identified similarity of score S is the probability that a score of at least S would have been obtained in matching by chance two unrelated sequences
- ✓ Very low values of the *p value* correspond to significant matchings (actually adopted threshold may go as low as 0.01%)

93

Problems and techniques I: Sequence data

Simple alignments

- ✓ The *E value* is related to the *p value* and is defined as the frequency of scores of at least S
- ✓ Implemented algorithms use sometimes slightly different approaches to computing the *p* and *E* values

94

Problems and techniques I: Sequence data

Simple alignments

- ✓ Divergent evolution may change protein sequences beyond recognition while preserving structure and function
- ✓ Examples are the family of globins, that share the same function (oxygen transport) and have the same structure and yet there are pairs of globins from different organisms whose sequences have less than 10% identical residues

95

Problems and techniques I: Sequence data

Simple alignments

- ✓ Detecting distant evolution relationships in biosequences is therefore important for biologists
- ✓ Unfortunately, algorithms like BLAST usually fail in recognizing such distant evolution relationships

96

Problems and techniques I: Sequence data

Simple alignments

- ✓ A way to go is to iteratively search for similar sequences, at each step using one (or more) sequence results of the previous step as the query sequence
- ✓ This is the simple idea implemented in the Psi-BLAST algorithm

97

Problems and techniques I: Sequence data

Simple alignments

- ✓ The algorithm works as follows
 - An initial BLAST search is first performed
 - From the second iteration on, Psi-BLAST computes a *sequence profile* summarizing the results of the previous iteration and uses it as the query sequence
- ✓ Psi-BLAST usually detects as much as twice as many significant evolutionary relationships as BLAST, and in some cases much more!

98

Problems and techniques I: Sequence data

Simple alignments

- ✓ The following table shows an example result obtained using Psi_BLAST on a quite common protein domain (Pleckstrin Homology) with quite divergent sequences (threshold E value set to 0.01) – computation terminated at the 4th iteration

Simple BLAST run

ITERATION NUMBER	NUMBER OF PH DOMAINS FOUND
1	93
2	607
3	622
4	622

99

Problems and techniques I: Sequence data

Multiple alignments

- ✓ Multiple alignment illustrates relationships holding amongst two or more sequences; for diverse sequences, common subsequences often correspond to key aminoacids associated with the maintenance of structural stability or function
- ✓ One of the most widely used package for multiple alignment is ClustalX at

www.hgmp.mrc.ac.uk/Registered/Option/clustalx.html

100

Problems and techniques I: Sequence data

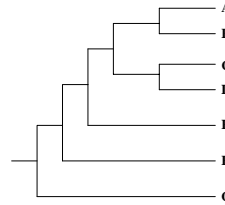
Multiple alignments

- ✓ As for the techniques that are usually applied, most approaches use the method of *progressive alignment*, which is very efficient but tends to preserve errors through iterations
- ✓ This amounts to construct a preliminary pairwise alignment, using its result to form a guide-tree and proceeding in accordance with the guide-tree to add sequences to the alignment one after another, starting with merging the most closely related and finishing with the most distant

101

Problems and techniques I: Sequence data

Multiple alignments



Guide- tree

Corresponding alignment order

1. A with B → AB
2. C with D → CD
3. AB with CD → ABCD
4. ABCD with E → ABCDE
5. ABCDE with F → ABCDEF
6. ABCDEF with G → ABCDEFG

102

Problems and techniques I: Sequence data

Multiple alignments

- ✓ Multiple alignments serve several purposes. An important one is to *assign proteins to families*, for a protein being assigned to a family tells much about its function
- ✓ Results of multiple alignments can be *condensed* in summary structures and stored in secondary databases, like the PROSITE database (www.expasy.org/prosite)
- ✓ There are several possibilities in order to summarize the information resulting from multiple alignments, including the alignment itself, consensus sequences, conserved residues, sequence patterns and so on

103

Problems and techniques I: Sequence data

Multiple alignments

- ✓ To construct *consensus sequences* amounts to produce a single sequence from a multiple alignment in which each residue is the most common (or consensus) for the analyzed sequence family
- ✓ Major weaknesses of this model are
 - information pertaining sequences that do not contain the consensus residue remains ignored, even if they may be useful
 - the consensus may be biased by the number of sequences as compared to the consensus threshold

104

Problems and techniques I: Sequence data

Multiple alignments

THRB_HUMAN	L	E	S	Y	I	D	G
THRB_BOVIN	F	E	S	Y	I	E	G
THRB_MOUSE	L	D	S	Y	I	D	G
THRB_RAT	L	D	S	Y	I	D	G
FA9_RAT	E	P	I	N	D	F	T
FA9_RABBIT	Q	S	S	D	D	F	T
CONSENSUS	X	X	S	Y	I	X	G

Serine protease sequences. Consensus threshold: 60%.

105

Problems and techniques I: Sequence data

Multiple alignments

- ✓ An alternative model consists in using *sequence patterns* that describe conserved regions using alternative symbol sequences
- ✓ For instance, the following two patterns maybe associated to the serine protease protein family:
 - [LIVM] – [ST] – A – [STAG] – H – C
 - [DNSTAGC] – [GSTAPIMVQH] – x(2) – G – [DE] – S – G – [GS] – [SAPHV] – [LIVMFYWH] – PA – [LIVMFYSTANQH]
- ✓ Square brackets group alternative symbols; x(n) represent *n* residues of any type

106

Problems and techniques I: Sequence data

Multiple alignments

- ✓ A weakness of this model is that it tends to be too much compact, thus leading to false positive occurrences in unrelated sequences
- ✓ A second problem is that it does not allow to attach probabilities to alternative symbols, which would be much useful in many cases
- ✓ For instance, PROSITE uses this model, but “corrects” entries with domain information

107

Problems and techniques I: Sequence data

Pattern extraction

- ✓ In biological applications, it is often useful to be able to delimit parts of sequences that are biologically meaningful
- ✓ Typical examples are determining the loci of genes, exons/introns boundaries in RNA and to detect unusual or over-represented subsequences
- ✓ Things are complicated by:
 - ✓ The size of involved sequences
 - ✓ The quite frequent occurrence of evolution-determined mutations

108

Problems and techniques I: Sequence data

Pattern extraction

- ✓ Quite diverse approaches have been developed to deal with pattern extraction problems in biosequences, including neural nets, data mining, grammars and automata, string indexing, HMMs and other statistical techniques
- ✓ In general, approaches to patterns extraction can be classified as either bottom-up or top-down and are either exact or approximate

109

Problems and techniques I: Sequence data

Pattern extraction

- ✓ In bottom-up approaches the basic idea is that of enumerate candidate patterns and computing some "fitness" measure in order to filter out uninteresting ones or for ranking them
- ✓ In top-down approaches models of the patterns are generated (either following some principle characteristics or inductively with techniques that may resemble multiple alignment); models are then validated using fitness measures

110

Problems and techniques I: Sequence data

Pattern extraction

- ✓ As far as top-down approaches are concerned, a potential way to go appears to be that of using grammars and automata in order to manage sequences as string to be parsed according to some rules, where such rules would be to define those subsequences one is interested in

111

Problems and techniques I: Sequence data

Pattern extraction

- ✓ For instance, palindromes and repetitions of groups of symbols often occur in those subsequences that are biologically meaningful (e.g., the so called GC islands, are usually associated to regions where genes are located)
- ✓ The bad news is that those regularities are actually significant from the biological point of view only if they occur in some specific contexts, and this makes the problems we are talking about highly ambiguous from the language-theoretic viewpoint

112

Problems and techniques I: Sequence data

Pattern extraction

- ✓ As an example, if we would build a grammar defining a gene, we could define something like the following:

$G \rightarrow PR$ P: promoter,
 $P \rightarrow N$ R: alternating exons and introns
 $R \rightarrow EIR \mid E$ E: exon; I: intron; N: nucleotide sequence
 $E \rightarrow N$ {a, c, g, t}: nucleotides
 $I \rightarrow gtNag$
 $N \rightarrow aN \mid cN \mid gN \mid tN \mid \epsilon$

- ✓ So, in this simple grammar, "gt" and "ag" are assumed to delimitate introns

113

Problems and techniques I: Sequence data

Pattern extraction

- ✓ That grammar is highly ambiguous for "gt" and "ag" may appear anywhere within an exon, an intron or in a promoter region
- ✓ Even if we were to introduce constraints about the nature of promoters, require that some length bounds on the size of exons and introns are met and so on, ambiguity would in any case remain
- ✓ This is determined by the very nature of biosequences; for instance, because of alternating splicing phenomena, the alternation exon/intron should be interpreted in different ways in different contexts

114

Problems and techniques I: Sequence data

Pattern extraction

- ✓ Thus, probabilities have to be introduced in the picture somehow
- ✓ A convenient formalism to do that is that of Hidden Markov Models (HMM), which can be seen as a variant of probabilistic finite-state transducers
- ✓ HMM can be profitably adopted, for instance, in order to represent protein domain families; this amounts to recognize, in a given protein sequence, those typical traits that make it belong to a protein family

115

Problems and techniques I: Sequence data

Pattern extraction

- ✓ An HMM comprises *match*, *delete* and *insert* states; states are linked by arrows denoting possible transitions between states; probabilities are associated to transitions and to states
- ✓ The model can be viewed as a way to generate set of sequences

116

Problems and techniques I: Sequence data

Pattern extraction

- ✓ For instance, in the case of protein sequences, a match state generate an aminoacid out of the possible 20 according to a probability distribution; different probability distributions apply to different match states and transitions, so contextual information can be taken into account
- ✓ A protein sequence is thus generated by "moving" through the model states; the corresponding path has a probability associated with it (computed from the probabilities associated to states and links in the path)

117

Problems and techniques I: Sequence data

Pattern extraction

- ✓ A family of proteins is represented by an HMM inasmuch as proteins belonging to that family are generated by the HMM with a very high probability
- ✓ Also, algorithms exist to evaluate the most probable path through the model to generate a given protein sequence, which is decided to belong to the family if the associated probability is above a given threshold

118

Problems and techniques I: Sequence data

Pattern extraction

- ✓ As for bottom-up approaches I am going to present an example technique in some detail shortly
- ✓ Before that I'd like to highlight the trade-off between efficiency and flexibility: methods specifically tailored to extract a certain class of patterns can be made quite effective and efficient, but are not applicable outside their specific domain
- ✓ This latter situation is indeed significant in some cases where even the general form of patterns to be extracted is fairly unknown

119

Problems and techniques I: Sequence data

An example technique

- ✓ A Frequent structured pattern is an example of particular class of repetitions that is biologically meaningful
- ✓ Frequent pattern occurrences are indeed associated with subsequences that encode, e.g., promoter regions and regulatory regions and are found in the context of certain pathologies

120

Problems and techniques I: Sequence data

An example technique

```
AGTGCAC TTTATATAGGA
TTACCGAGT CGCATACGT
CGAAGTATGATCTCGACC
AGTACCGGATCGACCTGCA
TGACTTAGTAACATACCA
TACCGAAGTACAGATTCGA
```

- ✓ *Structured:*
 - Two (or more) conserved regions spaced by constrained spacers

.....

- ✓ *Frequent:*
 - They are present at least q times (q is the *quorum*)

AGTA ???? TCG

Conserved regions Constrained spacer

121

Problems and techniques I: Sequence data

An example technique

- ✓ The problem of extracting frequent structured patterns can be defined in several ways, depending on:
 - the kinds of repetitions of conserved regions
 - exact
 - errors allowed
 - the kinds of structures
 - two conserved regions
 - multiple conserved regions
 - other variations which I shall address later

122

Problems and techniques I: Sequence data

An example technique

- ✓ Some definitions:
 - SC a set of strings
 - w a string pattern
 - w is an *exact motif* if it is repeated in at least q strings of SC
 - w' is an *e-occurrence* of w if $dis(w, w') \leq e$, where dis is a suitable distance function (for instance, Hamming or Levenstein distances can be used, which are relevant from the biological point of view)
 - w is an *e-motif* if it has an e-occurrence in at least q strings of SC

123

Problems and techniques I: Sequence data

An example technique

- ✓ definitions ...(continued)
 - Therefore, a structured pattern is a pattern of the form

$$p = w_{1[k_1]} X(d_1) w_{2[k_2]} \dots X(d_{r-1}) w_{r[k_r]}$$

where

$w_1[k_i]$ is a word of length k_i

$X(d_i)$ are d_i "don't care" symbols

124

Problems and techniques I: Sequence data

An example technique

$$p = \text{ACTX(4)GC}$$

- ✓ Occurrences of p in SC are:

CGA**ACT**TGATGCACC
 G**ACT**CCGGGCACCTGCA } **ACT????GC**

- ✓ Allowing e -occurrences in sequence repetitions means allowing for at most e errors in each word-box (Hamming distance)

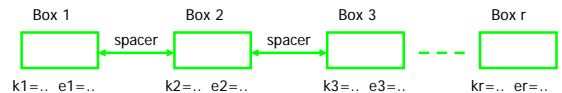
CGA**AGG**TGATGCACC 2 errors
 G**TCT**CCGGGCACCTGCA 1 error

125

Problems and techniques I: Sequence data

An example technique

- ✓ In general, we may look at subsequences we are interested in as shown in the figure below:



126

Problems and techniques I: Sequence data

An example technique

- ✓ A structured pattern p is a *structured exact motif* if there are at least q strings in SC containing an occurrence of p
- ✓ A structured pattern p is a *structured e -motif* if there are at least q strings in SC containing an e -occurrence of p .

127

Problems and techniques I: Sequence data

An example technique

- ✓ Problems definition:
 - Problem 1: Find all *structured exact motifs* of the form $m = w_{1[k_1]}X(d)w_{2[k_2]}$
 - Problem 2: Find all *structured exact motifs* of the form $m = w_{1[k_1]}X(d_1)w_{2[k_2]} \dots X(d_{r-1})w_{r[k_r]}$
 - Problem 3: Find all *structured e -motifs* of the form $m = w_{1[k_1]}X(d)w_{2[k_2]}$
 - Problem 4: Find all *structured e -motifs* of the form $m = w_{1[k_1]}X(d_1)w_{2[k_2]} \dots X(d_{r-1})w_{r[k_r]}$

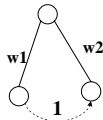
128

Problems and techniques I: Sequence data

An example technique - Problem 1

4. Create the d-link between the inserted words or update the number of occurrences of the associated pattern
5. Note that either w1 or w2 might be already present in the tree

w1 w2
AGTGCACTTTATATAGGA
TTACCGAGTCGCATACGT
CGAAGTATGATCTCGACC
AGTACCGGATCGACCTGCA
TGACTTAGTAACATACCA



133

Problems and techniques I: Sequence data

An example technique - Problem 1

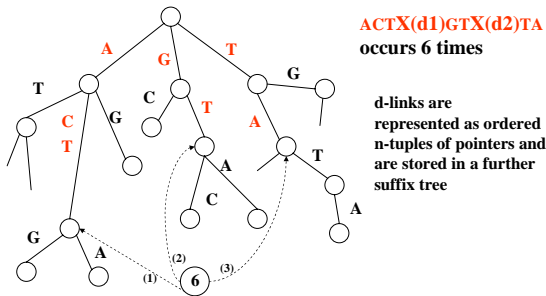
- ✓ When the number of occurrences of a d-link equals the quorum q , it is recognized as a structured motif
- ✓ Therefore structured motifs are recognized during the construction of the tree (this saves time)
- ✓ The complexity figures we obtain are ($n = |SC|$):
 - $O(n)$ time
 - $O(n)$ space

134

Problems and techniques I: Sequence data

An example technique - Problem 2

It suffices to extend d-links to more than 2 nodes



135

Problems and techniques I: Sequence data

An example technique

- ✓ Problems definition:
 - Problem 1: Find all *structured exact motifs* of the form $m = w_{1[k_1]}X(d)w_{2[k_2]}$
 - Problem 2: Find all *structured exact motifs* of the form $m = w_{1[k_1]}X(d_1)w_{2[k_2]} \dots X(d_{r-1})w_{r[k_r]}$
 - Problem 3: Find all *structured e-motifs* of the form $m = w_{1[k_1]}X(d)w_{2[k_2]}$
 - Problem 4: Find all *structured e-motifs* of the form $m = w_{1[k_1]}X(d_1)w_{2[k_2]} \dots X(d_{r-1})w_{r[k_r]}$

136

Problems and techniques I: Sequence data

An example technique

- ✓ Allowing mismatches in word-boxes is highly significant from the biological viewpoint
- ✓ Several ways of measuring "mismatches" can be adopted
- ✓ Two already mentioned ones are
 - the Hamming distance $h(w_1, w_2)$: the minimum number of substitutions for obtaining w_2 from w_1
 - the Levenstein distance $l(w_1, w_2)$: the minimum number of edit operation needed for obtaining w_2 from w_1
- ✓ Clear enough, Hamming requires equal length

137

Problems and techniques I: Sequence data

An example technique

- ✓ Let us consider the Hamming distance
- ✓ Given a word w of length k , an alphabet Σ and a maximum number of mismatches e , there are $\sum_{i=0}^e \binom{k}{i} (|\Sigma|-1)^i$ possible words w_j s.t. $h(w, w_j) \leq e$
- ✓ The problem is to represent this set of words in a way as compact as possible

138

Problems and techniques I: Sequence data

An example technique

- ✓ The idea to solve the problem is as follows:
 - Consider the maximum allowed number of errors
 - Substitute mismatching symbols with "don't care" symbols (X)
 - Consider the words (after the substitution) having exactly e don't care symbols
- ✓ There are $\binom{k}{e}$ strings of this sort thus generated (e-neighbors)

139

Problems and techniques I: Sequence data

An example technique

- ✓ As an example, consider the string $w=AGCT$ and let $e=2$. Then the e-neighbor set is
 $\{AGXX, AXCX, AXXT, XG CX, XGXT, XXCT\}$
- ✓ By "instantiating" don't care symbols over the alphabet, we obtain all the words w_j such that $h(w, w_j) \leq e$

140

Problems and techniques I: Sequence data

An example technique – Problems 3 and 4

In order to solve Problem 3, the following can be done:

1. For each structured pattern in SC derive its set of e-neighbor patterns
2. Count the occurrences of e-neighbor patterns
3. Exploit the technique for Problem 1 considering don't care symbols as symbols in the alphabet
4. For each structured pattern in SC combine the number of occurrences of its e-neighbor patterns

141

Problems and techniques I: Sequence data

An example technique – Problems 3 and 4

- ✓ It should be noted that, in order to count occurrences, it is not sufficient to simply sum up the number of occurrences of e-neighbor patterns (the set of strings in which two e-neighbor patterns have been found may overlap)
- ✓ To solve this latter problems, it is sufficient to store, for each e-neighbor pattern, a boolean array indicating in which string it has been found

142

Problems and techniques I: Sequence data

An example technique – Problems 3 and 4

- ✓ Problem 4 is finally solved starting from the solution of Problem 3 in a way similar as Problem 2 is solved generalizing the solution for Problem 1

143

Problems and techniques I: Sequence data

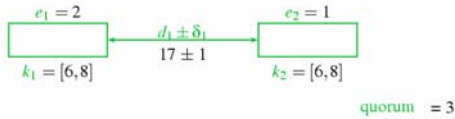
An example technique – The flexibility issue

- ✓ However, especially in eukaryotes, it is often not known which classes of patterns are indeed interesting
- ✓ Therefore it would be interesting to allow variable pattern formats to be specifiable and resolved

144

Problems and techniques I: Sequence data

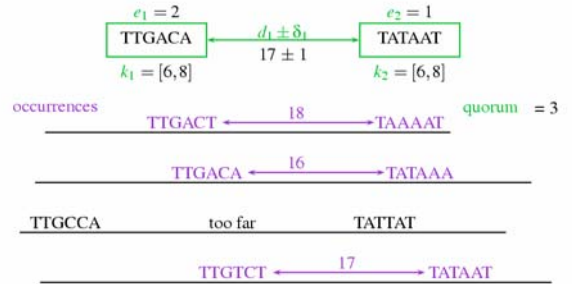
An example technique - The flexibility issue



145

Problems and techniques I: Sequence data

An example technique - The flexibility issue



146

Problems and techniques I: Sequence data

An example technique - The flexibility issue

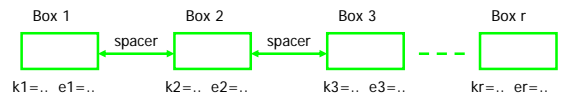
- ✓ The example technique discussed before can be embedded within an efficient problem-generation machinery in order to attain high flexibility at the cost of a reasonable performance degradation

147

Problems and techniques I: Sequence data

An example technique - The flexibility issue

- ✓ The general structure of supported pattern extraction problems is:



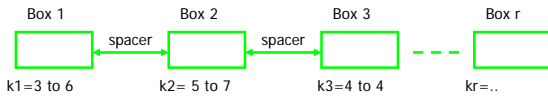
148

Problems and techniques I: Sequence data

An example technique - The flexibility issue

Where the following variants can be specified:

- Box number:
 - Fixed (e.g. find patterns with 3 boxes)
 - Variable within an interval (e.g., any pattern with at least 3 and at most 6 boxes)
- Box length:
 - All boxes with the same length
 - Individual box lengths
 - Individual box lengths each varying within an interval
 - Any (admissible lengths are computed off-line)



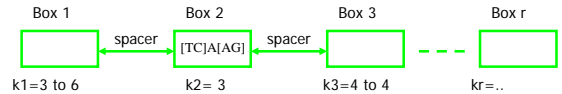
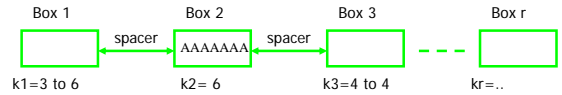
149

Problems and techniques I: Sequence data

An example technique - The flexibility issue

Box content:

- Specifying "anchor" boxes having a specific content
- Using the formalism of sequence pattern



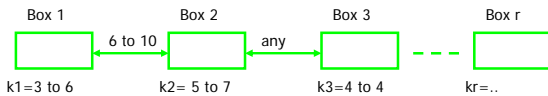
150

Problems and techniques I: Sequence data

An example technique - The flexibility issue

Box distances:

- All boxes at the same distance
- Box i at distance d_i from Box $i+1$ (for each i) - exact distance
- Variable distances within an interval
- Any (admissible distances are computed off-line)



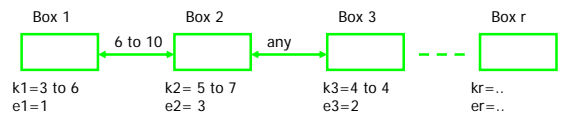
151

Problems and techniques I: Sequence data

An example technique - The flexibility issue

Similarity across patterns

- Exact match
- Hamming Distance
 - maximum allowed distance required can be different for different boxes
 - equal index boxes must have the same lengths
- Levenstein Distance
 - maximum allowed distance required can be different for different boxes
 - equal index boxes might have different lengths



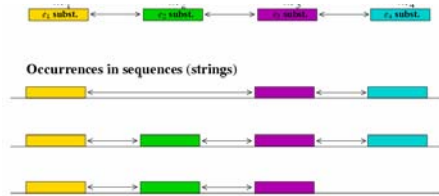
152

Problems and techniques I: Sequence data

An example technique – The flexibility issue

✓ Special repetition options

- Allow skips

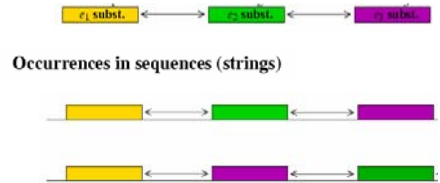


153

Problems and techniques I: Sequence data

An example technique – The flexibility issue

- Allow order changings



154

Problems and techniques I: Sequence data

An example technique – The flexibility issue

- Allow **single box inverse**
 - ACGT-----TTAGC----GGACC
 - ACGT-----CGATT----GGACC
 - Can be a correct match
 - Similarity options (Exact, Hamming, Levenstein) are applied also to inverted boxes
- Allow **inverse patterns**
 - ACGT-----TTAGC----GGACC
 - CCAGG-----CGATT----TGCA

155

Outline

Along the presentation, I will go through the following main points:

- Introduction
- Some basic molecular biology
- Biological datasets formats and repositories
- Research goals in bioinformatics
- Overview of problems and techniques I: Sequence data
 - Simple and multiple alignments
 - Pattern extraction from sequences
 - An example pattern extraction technique
- Overview of problems and techniques II: Structural data
 - Determining macromolecule structure
 - Protein structure prediction
 - RNA structure prediction
- Some further issues
- Conclusion

156

Problems and techniques II: Structural data

Determining macromolecule structure

- ✓ I have already mentioned several times that 3D structure taken by macromolecules in their environment is fundamental in understanding their function
- ✓ DNA has a known shape
- ✓ Determining protein structure is of central importance in biology
- ✓ Also interesting is to establish the geometrical structure assumed by RNA sequences

157

Problems and techniques II: Structural data

Determining protein structure

- ✓ Two are the principle lab methods used to single out the 3D shape of a protein: X-ray crystallography and NMR spectroscopy
- ✓ Both methods involve long times and significant costs
- ✓ Alternative Sw methods are those that *predict* proteins' structures on the basis of homology-driven strategies or using chemical or physical characterizations of the involved residues (ab-initio methods)

158

Problems and techniques II: Structural data

Determining protein structure

- ✓ X-ray crystallography involves the determination of protein structure by studying the diffraction pattern of X-rays through a precisely orientated protein crystal. The way in which X-rays are scattered depends on the electron density and spatial orientation of atoms in the crystal
- ✓ The method is quite difficult to apply to those proteins that do not easily form crystals

159

Problems and techniques II: Structural data

Determining protein structure

- ✓ NMR uses the property of atoms to switch between magnetic states in an applied magnetic field. The nature of the phenomenon depends by specific atoms and chemical context, so that NMR spectroscopy can discriminate between chemical substances
- ✓ The technique can be applied to small soluble proteins

160

Problems and techniques II: Structural data

Protein structure prediction

- ✓ Since lab methods require significant amounts of time, data about protein structures are gathered at a much lower rate as compared to sequence data
- ✓ It is therefore quite interesting to have available methods capable of *predicting* protein structure on the basis of their aminoacid sequence
- ✓ Protein structure prediction methods falls in several categories; I will discuss ab-initio and knowledge-based protein prediction as well as team prediction

161

Problems and techniques II: Structural data

Protein structure prediction

- ✓ The possibility to predict protein structures *computing* them stems from the consideration that proteins can fold to their native structures spontaneously, without the intervention of any other agent: this means that the protein fold is indeed encoded in its aminoacid sequence
- ✓ It is also important to keep in mind that even relatively small proteins are associated with a very large number of possible structures, so that the determination of the right one cannot proceed by exhaustive search

162

Problems and techniques II: Structural data

Protein structure prediction

- ✓ In general, predicting a protein 3D structure means predicting the relative position of every atom occurring in the protein solely on the basis of the protein sequence
- ✓ A way to go is to consider that proteins fold into a shape corresponding to a minimal level of free energy
- ✓ Therefore, it is in principle possible to predict the 3D structure of the protein at hand as the one minimizing such energy
- ✓ This is, by the way, very difficult for the systems modeling proteins are huge

163

Problems and techniques II: Structural data

Protein structure prediction

- ✓ Ab-initio methods are quite attractive from the intellectual viewpoint, but much has to be done in general to improve their prediction quality
- ✓ Several such methods have been however developed, one notable example being the Rosetta predictor, available at

www.bioinfo.rpi.edu/~bystrc/hmmstr/server.php

164

Problems and techniques II: Structural data

Protein structure prediction

- ✓ Alternatively, knowledge-based methods attempt to predict protein structure using relevant information stored in available databases
- ✓ The idea here is that similar sequences will correspond to similar 3D structures
- ✓ This has a theoretical basis in Sander and Schneider's principle that sequences showing more than 25% identity over an alignment of 80 residues adopt the same structure; there are exceptions, though

165

Problems and techniques II: Structural data

Protein structure prediction

- ✓ Comparative modeling and fold recognition methods fall in this category
- ✓ Knowledge-based methods seem, at the moment, to guarantee better accuracy than ab-initio techniques
- ✓ It is worth mentioning the CASP (Critical Assessment of Structure Prediction) context where protein structure predictors participate in a competition to evaluate them comparatively; CASP contexts are illustrated at

<http://predictioncenter.llnl.gov/>

166

Problems and techniques II: Structural data

Protein structure prediction

- ✓ Structure prediction by comparative (or homology) modeling works by computing an alignment complying with Sander and Schneider's threshold between the given target sequence (the protein whose 3D shape is to be predicted) and one or more sequences (templates) whose structure is known
- ✓ Lack of suitable templates obviously limits the applicability of this technique
- ✓ Good alignments usually produce good prediction, but experts' verification is always in order

167

Problems and techniques II: Structural data

Protein structure prediction

- ✓ Prediction computation proceeds by first looking for the template structure(s); if more than one is found, an average structure is computed by averaging atom positions
- ✓ Using the alignment of the target sequence to the template, the structure is divided in regions: the structural core, the loops and the side chains
- ✓ Structure prediction of the core is easy, since atom positions in the backbone of the averaged templates are predicted as the backbone atoms in the target structure

168

Problems and techniques II: Structural data

Protein structure prediction

- ✓ Loops are more difficult to handle
- ✓ The simplest method to deal with them is the *spare-parts* algorithm, that uses a database of known loop structures (spare loops) from other proteins (not necessarily similar to the target one)
- ✓ For each loop to be predicted, the spare loop that best fit in the gap of the modeled structure is chosen and its 3D structure is returned as the 3D structure predicted for the given loop
- ✓ A similar approach is used for side-chains

169

Problems and techniques II: Structural data

Protein structure prediction

- ✓ In order to improve the prediction it is sometimes used to verify energy-minimization constraints by which predicted atom positions may be sometimes slightly modified
- ✓ A comparative modeling tool is available at

<http://swissmodel.expasy.org/>

170

Problems and techniques II: Structural data

Protein structure prediction

- ✓ Comparative modeling cannot be applied when no suitable template structure exists
- ✓ In such cases, secondary structure prediction can be a viable alternative
- ✓ It should be noted that it does not produce a full atom model of the tertiary structure but rather it provides a prediction of the secondary structure state of each residue (either helical, strand or coil)

171

Problems and techniques II: Structural data

Protein structure prediction

- ✓ Most methods for secondary structure prediction are trained on databases of known structures
- ✓ Some approaches use the concept of *secondary structure propensity*, that is, aminoacids seem to prefer certain secondary structures states
- ✓ However, propensity is not enough, since this is never truly strong: other considerations are applied as well

172

Problems and techniques II: Structural data

Protein structure prediction

- ✓ For instance, in the Chou-Fasman method, an helix is predicted only if there are at least 4 helix-favouring residue in a run of 6; the helix is extended along the sequence until a proline is encountered (which is known to break helices) or a run of 4 residues with low helical propensity is found; similar rules apply to strand predictions
- ✓ Those methods are not so accurate though
- ✓ Multiple sequence alignments to known structures can be used to significantly improve the prediction

173

Problems and techniques II: Structural data

Protein structure prediction

- ✓ Also, machine learning methods and neural networks have been successfully developed
- ✓ Those methods are capable of correctly predicting most of the protein secondary structure, errors being made above all on residues that lie at the end of each secondary structure
- ✓ PSI-PRED is one such predictors; can be found at

<http://bioinf.cs.ucl.ac.uk/psipred/>

174

Problems and techniques II: Structural data

Protein structure prediction

- ✓ For proteins, similarity in sequences is usually sufficient to establish similarity in structure
- ✓ However, evolution has conserved structure much more than sequences
- ✓ There are cases of distantly related protein with the same structure but whose level of sequence similarity is well below the 25% of Sander and Schneider's principle
- ✓ Fold recognition techniques aim at detecting such structural similarities

175

Problems and techniques II: Structural data

Protein structure prediction

- ✓ Fold recognition can be seen as an extension of the comparative modeling method to very distant relationships
- ✓ They operate by searching through a library of known protein structures (fold library) and finding the one most compatible with the target sequence whose structure has to be predicted
- ✓ Then, the resulting alignment is used to construct the prediction

176

Problems and techniques II: Structural data

Protein structure prediction

- ✓ Fold recognition methods use a variety of approaches, often mixing (even limited) similarity detected in sequences using substitution matrices with structural information
- ✓ An example of such methods are 3D-PSSM and its follow-up called Phyre; those packages are found at

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

177

Problems and techniques II: Structural data

Protein structure prediction

- ✓ Before closing with this part of the tutorial, I am going to briefly describe a protein prediction strategy based on the idea of predicting by team-working
- ✓ This is an example of how database-like technique can be applied, even indirectly, to bioinformatics problems to substantiate viable solutions
- ✓ The database competences applied in this case lie in the area of database integration

178

Problems and techniques II: Structural data

Protein structure prediction

- ✓ Some approaches propose Meta-Predictors, i.e. they query several tools on the same problem and choose the “best” prediction among the returned ones
- ✓ However, meta-prediction does not allow to use the fact that, for the same protein sequence, there are predictors that compute best results on some portions of the sequence, while other predictors work better over other parts of it
- ✓ The idea is therefore to find a way to combine the results of several tools to improve the overall result quality

179

Problems and techniques II: Structural data

Protein structure prediction

To do that, several problems have to be solved:

- Inputs and outputs of various tools usually have diverse formats: a uniform representation is needed
- It is unfeasible to integrate various techniques at the algorithmic level (they may have quite diverse nature, say, ab-initio or homology-based)
- Difficulty to integrate the outputs because they compute an otherwise unknown structures; How to choose the “best” parts?
- Generally, available tools are specialized for working well with some families of proteins, whereas a general method would be useful that adapts to (almost) all situations

180

Problems and techniques II: Structural data

Protein structure prediction

- ✓ Evaluation of the quality of a prediction tool - intuitively:
 - Given a protein p and the prediction of the tool
 - How much the prediction approximates the real structure?
 - Possible to be computed only for proteins whose structure is known!!
- ✓ Definition of Application Domain
 - A set of proteins whose structure is known and having characteristics similar to the ones we are interested to predict

181

Problems and techniques II: Structural data

Protein structure prediction

- ✓ Given an application domain D
 - Evaluate the performances on D of the available tools when they are applied singularly (by using proteins whose structure is known)
 - Evaluate the performances of a team of predictors when they are jointly applied on D

182

Problems and techniques II: Structural data

Protein structure prediction

- ✓ Select the team T best performing on D
- ✓ Define integration rules for the results yielded by different tools to obtain a single prediction
- ✓ Apply T on proteins whose structure is unknown and integrate the results to obtain the final overall prediction

183

Problems and techniques II: Structural data

Protein structure prediction

- ✓ Evaluation of the performances of a single tool PT
 - Given a known protein p , its true three dimensional structure $tdp0$ and the prediction $tdpk$ of PT for p align their three dimensional structures
 - Perform a pairwise comparison of the three dimensional coordinates
 - Measure the Overlap Degree between $tdp0$ and $tdpk$
 - Average the Overlap Degrees over the proteins of D to measure the precision of PT over D

184

Problems and techniques II: Structural data

Protein structure prediction

- ✓ To select the best team:
 - Order tools by decreasing precision
 - A tool T_k is added to the team if the team precision computed over D increases by adding T_k
 - Stop when no new tool addition improves the team precision
- ✓ To compute the overall prediction, use a biased voting algorithm, where biases correspond to the reliability of predictors

185

Problems and techniques II: Structural data

Protein structure prediction

- ✓ Experiments.....
- ✓ Five predictors considered
- ✓ Several proteins whose structure is known (to assess the performances)
- ✓ Use of SCOP database to determine Domains
- ✓ The method always improves (usually significantly!!) over the best prediction

186

Problems and techniques II: Structural data

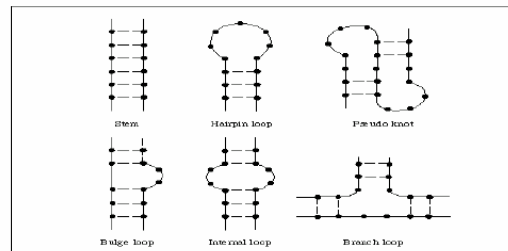
RNA structure prediction

- ✓ RNA is transcribed in cells as a single strand of nucleotides. These strands fold into structures because of the interactions between nucleotides
- ✓ Basic RNA geometrical structures are different than those of proteins and comprise hairpins, loops, stems, bulges, knots

187

Problems and techniques II: Structural data

RNA structure prediction



188

Problems and techniques II: Structural data

RNA structure prediction



An image of a folded RNA strand

189

Problems and techniques II: Structural data

RNA structure prediction

- ✓ Several methods are used to predict the RNA structure. From the primary structure the secondary structure can be predicted since this can expose biologically relevant features
- ✓ On the other hand, RNA tertiary structure is very difficult to predict and, in fact, tertiary structure of RNA is investigated by X-ray crystallography and NMR

190

Problems and techniques II: Structural data

RNA structure prediction

- ✓ As with proteins, however, secondary structure predictions can be used along with NMR and X-ray crystallography in determining 3D structure
- ✓ There are two main approaches used to predict the secondary structure, that are, energy minimization and comparative sequence analysis
- ✓ Most available methods use free energy estimates. These methods look for the fold with the lowest free energy out of possible folds.

191

Problems and techniques II: Structural data

RNA structure prediction

- ✓ Similarly to proteins, comparative analysis uses multiple sequence alignments of homologous sequences to predict the structure. This usually needs many aligned sequences.

192

Problems and techniques II: Structural data

Protein structure prediction

Along the presentation, I will go through the following main points:

1. Introduction
2. Some basic molecular biology
3. Biological datasets formats and repositories
4. Research goals in bioinformatics
5. Overview of problems and techniques I: Sequence data
 - a. Simple and multiple alignments
 - b. Pattern extraction from sequences
 - c. An example pattern extraction technique
6. Overview of problems and techniques II: Structural data
 - a. Determining protein structure
 - b. Protein structure prediction
 - c. RNA structure prediction
7. **Some further issues**
8. Conclusion

193

Some further issues

- ✓ There are several issues which I have not discussed in this tutorial but relevant to bioinformatics
- ✓ The following slides shortly address some of them, that are:
 - Phylogenetics
 - Microarray data analysis
 - Proteomic data analysis
 - Molecular pathways
 - Bioinformatics and drug synthesis
 - Visualization tools

194

Some further issues - Phylogenetics

- ✓ Evolutionary relationships among species are important because similar species are characterized by similar mechanisms at the macromolecular level (notably, proteins)
- ✓ Phylogenesis aims at reconstructing such evolutionary relationships
- ✓ Most widely used technique is the analysis of biosequences via clustering
- ✓ Understanding the role and the pace of mutations occurring in sequences is central to reconstruct correct evolutionary scenarios

195

Some further issues - Phylogenetics

- ✓ Phylogenetic trees show evolutionary relationships, where nodes represent species and links stand for lines of descent
- ✓ The basic idea with using sequence analysis in this context is that DNA (and, hence, RNA and proteins) accumulate mutations over evolutionary time leading to divergence in different lines of descents

196

Some further issues - Phylogenetics

- ✓ Depending on the phylogeny to be studied, different macromolecules should be used
- ✓ For instance, mitochondrial DNA is a good pick for studying closely related organisms, since it evolves rapidly, whereas highly conserved macromolecules should be used for studying distant species
- ✓ Other techniques (such as protein folds recognition) can be adopted in studying distant evolutionary relationships

197

Some further issues - Phylogenetics

- ✓ Caution is needed in that it is difficult to guarantee that a given phylogenetic tree accurately represent evolutionary history
- ✓ Multiple-tries approach can be used to improve reliability
- ✓ Phylogenetics is obviously related to species classification systems and ontologies

198

Some further issues - Microarray analysis

- ✓ A microarray consists of a set of genetic elements (called features) arranged in a regular grid over a convenient substrate (e.g., a glass chip)
- ✓ Microarrays can be used to analyse the composition of genetic material in a tissue sample
- ✓ Genetic elements are single strand DNAs (probes)
- ✓ They use the fact that complementary nucleotides pair to one another

199

Some further issues - Microarray analysis

- ✓ Therefore, if a tissue sample macromolecule binds to a probe ATCGGC then it can be concluded that that macromolecule sequence is TAGCCG
- ✓ The same holds for RNA
- ✓ Nucleotide sequences expression levels are proportional to the level of protein expressed by a gene
- ✓ The relative expression levels of thousands of genes can be analysed this way on a single chip

200

Some further issues - Proteomic analysis

- ✓ Proteins expression in tissue and serum samples gives interesting insights to biologists
- ✓ Besides microarray techniques, 2D-page gels and mass spectrometry can be used to this end

201

Some further issues - Proteomic analysis

- ✓ 2D-page is a protein separation technique that allows the analysis of many proteins on a single gel
- ✓ Separated proteins appear as spots; the quantity, level and distribution of such spots within the gel translates into the proteomic profile of the give sample
- ✓ Data from 2D-page experiments can be accessed via Web, e.g., through the ExpASy site at www.expasy.ch/ch2d/2d-index.html

202

Some further issues - Proteomic analysis

- ✓ Data from mass spectrometry experiments are the mass/charge ratios of ions in a vacuum, which allow to determine peptide mass fingerprints (and, thus, protein expression levels, using protein database searching)
- ✓ Multiple Web sources are available, for instance: prospector.ucsf.edu

203

Some further issues - Molecular pathways

- ✓ Proteic and nucleotide sequences can be studied individually
- ✓ However, insights into their functions can be gained by studying (functional and structural) complex systems where they are embedded, like molecular pathways, tissues, organs and so on
- ✓ Molecular pathways can be represented by graphs, where nodes denote macromolecules and links denote relationships; for instance, in regulatory pathways, nodes represent protein and links represent information transfers between them

204

Some further issues – Molecular pathways

- ✓ Pathways can be reconstructed from protein expression data and protein interaction data
- ✓ Also, they can be modeled and simulated using differential equations
- ✓ A notable Web resource for pathways is
www.genome.ad.jp/kegg/

205

Some further issues – Drug synthesis

- ✓ Drugs interact with targets (usually, proteins) in the body and it is this interaction that causes responses
- ✓ Genomics and proteomics have contributed to massive increase of data available for drug design and testing
- ✓ Specific bioinformatic applications in this context include, for instance:
 - Modeling protein interaction with small molecules for drug design
 - Association of genotypes with drug response patterns
 - Processing and storing large data sets from testing campaigns

206

Some further issues – Visualization tools

- ✓ Several Sw exist for visualizing biological structures
- ✓ Some of them are listed below
 - RasMol www.umass.edu/microbio/rasmol
 - Cn3D <http://ncbi.nih.gov/Structure/CN3D/cn3d.html>
 - Chime accessible from the RasMol Web page
 - Molscript www.avatar.se/molscript
 - Ribbons sgce.cbse.uab.edu/ribbons
 - TOPS www.tops.leeds.ac.uk/~roman/surfnet/surfnet.html

207

Outline

Along the presentation, I will go through the following main points:

1. Introduction
2. Some basic molecular biology
3. Biological datasets formats and repositories
4. Research goals in bioinformatics
5. Overview of problems and techniques I: Sequence data
 - a. Simple and multiple alignments
 - b. Pattern extraction from sequences
 - c. An example pattern extraction technique
6. Overview of problems and techniques II: Structural data
 - a. Determining protein structure
 - b. Protein structure prediction
 - c. RNA structure prediction
7. Some further issues
8. **Conclusion**

208

Conclusion

- ✓ Bioinformatics is a challenging field where to apply computer science (and database!) competences
- ✓ Database-like techniques seem to be promising for several relevant application contexts (data mining, metadata management, source and service integration, incomplete information, data indexing,)
- ✓ Continuous collaboration with biologists is mandatory (despite cultural-dependent difficulties)

209

Conclusion

- ✓ I close this tutorial with some pointers to interesting journals/conferences dealing with bioinformatics:
 - Journals:
 - ═ Bioinformatics
 - ═ IEEE/ACM Trans. On Computational Biology and Bioinformatics
 - ═ Journal of Computational Biology
 - ═ Journal of Bioinformatics and Computational Biology
 - ═ Several biology journals like Science, Nature, Nucleic Acid Research, Journal of Molecular Biology

210

Conclusion

- Conferences:
 - ═ Research in Computational Molecular Biology
 - ═ IEEE Computer Society Bioinformatics Conference
 - ═ Pacific Symposium on Biocomputing
 - ═ Intelligent Systems for Molecular Biology
 - ═ And, by the way, many computer science conferences are opened to submission of papers in bioinformatics

211

Conclusion

- ✓ ... and listing next some of the main Web gateways to bioinformatics databases, tools and other material:

www.ncbi.nlm.nih.gov	National center for biotechnology information homepage
www.ebi.ac.uk	The EMBL European Bioinformatics Institute
www.expasy.ch	The Expasy (Expert protein analysis system) Molecular Biology Server
www.embl-heidelberg.de	European Molecular Biology Laboratory homepage
www.bmn.com	The BioMedNet gateway to biological sites
www.genomesonline.org	Genomes on line with links to genome databases and projects
www.genome.ad.jp/kegg	The Kyoto Encyclopedia of Genes and Genomes

212

Acknowledgments

I would like to thank:

- ✓ Enrico, Letizia and Maurizio and the other guys in SEBD organization for having allowed me the opportunity to give this tutorial
- ✓ MIUR, for partially funding my research in bioinformatics under the PRIN'03 program "Combinatorial and algorithmic methods for pattern discovery in biosequences"
- ✓ And, last but not the least, all of You for the patience you had in listening at me

213

That's all folks!

214