Using MODELLER for Comparative Modeling

D. R. Ripoll

August 11, 2002

- **MODELLER** is a program for comparative modeling written by Prof. Šali’s group at Rockefeller University.
- The program uses a **scripting language**.
- The **user provides an alignment** of a sequence to be modeled with known related structures.
- **MODELLER automatically calculates a model** with all non-hydrogen atoms.

CBSU

CTC

The **input** are:

- Protein Data Bank (PDB) atom files of known protein structures;
- their alignment with the target sequence to be modeled.

The **output** is a model for the target that includes all non-hydrogen atoms.

- MODELLER can calculate sequence and structure alignments, however, it is better to prepare the alignment carefully by other means.

There are three kinds of input files:

- **Atom files** (coordinates for the template structures);
- **Alignment file** (contains the alignment of the template structures with the target sequence);
- **Script file** (tells MODELLER what to do).

The Atom Files

Protein Data Bank atom files with coordinates for the template structures.

- Each atom file is named ***code*.atm** (or code.pdb) where ***code*** is a short protein code, preferably the PDB code; for example, *Peptococcus aerogenes* ferredoxin would be in a file **1fdx.atm**.
- The code **must** be used as that protein's identifier throughout the modeling.
- The atom sets do not have to be superposed by the user before comparative modeling is done.

Selecting the templates

- **Significantly different** structures homologous in sequence to the target one **should be used**.
- However, it **is not always best to use all related 3D structures** as templates because the objective function may become too rugged (six templates is a large number of templates).
- If two relatively similar templates have been solved at a high and low resolution; **use only the high resolution template**.
- Other factors to be considered in the selection of templates: **ligands bound** to the template and/or target; template structure from **NMR** or **X-Ray** crystallography.

The Alignment File

- One of the formats for the alignment file is related to the PIR database format; this is the preferred format for comparative modeling:

```
C; A sample alignment in the PIR format; used in tutorial
 >P1;5fd1
structureX:5fd1:1 : :106 : :ferredoxin:Azotobacter vinelandii: 1.90: 0.19
AFVVTDNCIKCKYTDCVEVCPVDCFYEGPNFLVIHPDECIDCALCEPECPAQAIFSED
EVPEDMQEFIQLNAELAEVWPNITEKKDPLPDAEDWDGVKGKLQHLER*
>P1;1fdx
sequence:1fdx:1 : :54 : :ferredoxin:Peptococcus aerogenes: 2.00:-1.00
AYVINDSC--IACGACKPECPVNIIQGS--IYAIDADSCIDCGSCASVCPVGAPNPED
------------------------------------------------*
```

The Script File

The script file contains commands for MODELLER, in the TOP language.

A sample script file that generates a model of sequence 1fdx from the structure of 5fd1

```
# Homology modelling by the MODELLER TOP routine 'model'.

INCLUDE                                          # Include the predefined TOP routines
SET OUTPUT_CONTROL = 1 1 1 1 1                   # uncomment to produce a large log file
# SET OUTPUT_CONTROL = 1 1 1 1 0                 # write real_output, notes, warnings, errors, dynmem
SET ALNFILE = 'alignment.ali'                    # alignment filename
SET KNOWNS = '5fd1'                              # codes of the templates
SET SEQUENCE = '1fdx'                            # code of the target
SET ATOM_FILES_DIRECTORY = './:../atom_files'    # directories for input atom files
SET STARTING_MODEL= 1                            # index of the first model
SET ENDING_MODEL = 1                             # index of the last model
                                                 # (determines how many
                                                 # models to calculate)
CALL ROUTINE = 'model'                           # do homology modelling
```

How is the alignment used by MODELLER?

- **For the aligned regions**, MODELLER tries to derive a **3D model for the target sequence that is as close to the template structure as possible**, while also satisfying stereochemical restraints (i.e., bond lengths, angles).

- **For the inserted regions,** which do not have any equivalent segments in any of the templates, are **modeled in the context of the whole molecule, but using their sequence alone.**

- Thus, when a user aligns a target residue with a template residue, he is telling MODELLER to **treat the aligned residues as structurally equivalent**.

Previous Example

```
>P1;5fd1
structureX:5fd1:1 : :106 : :ferredoxin:Azotobacter vinelandii: 1.90: 0.19
AFVVTDNCIKCKYTDCVEVCPVDCFYEGPNFLVIHPDECIDCALCEPECPAQAIFSED
EVPEDMQEFIQLNAELAEVWPNITEKKDPLPDAEDWDGVKGKLQHLER*
>P1;1fdx
sequence:1fdx:1 : :54 : :ferredoxin:Peptococcus aerogenes: 2.00:-1.00
AYVINDSC--IACGACKPECPVNIIQGS--IYAIDADSCIDCGSCASVCPVGAPNPED
------------------------------------------------*
```

Becoming familiar with the family fold and improving the alignment

Important stage to improve the model significantly:

- It mostly **involves visual inspection** of the superposed (experimental) structures on a graphics terminal.
- The aim is to study the family fold, to establish the relationships between various members of the family, and to **determine which regions are conserved and which are variable**.
- For example, **remove gaps from inside a-helical segments or b-strands;** gaps should be moved into those exposed regions showing large variations in the family of known structures and to the tips of loops.

Running MODELLER

To **run MODELLER** with the script file model-default.top, execute the following command:

C:\mod6v2\bin\mod.bat model-default.top 1>>model-default.log 2>>err

(or ***mymod.bat model-default***)

In Unix systems: ***mod model-default***

CB SU

CTC

Running MODELLER

- A number of intermediary files are created as the program proceeds. After a few seconds , the final **1fdx** model is written to file **1fdx.B99990001**.
- **Examine** the **model-default.log** file for information about the run.
- One should always check the output of the **CHECK_ALIGNMENT** command by searching for `**chkaln**'.
- Also, check for **warning** and **error messages** by searching for **`_W>'** and **`_E>',** respectively.
- **There should be no error messages**.
- **Warning** messages that can usually be ignored.

CBSU

CTC

Evaluation of the model

- Graphics: **Rasmol**, MOLMOL

  Check secondary structure elements; distribution of hydrophobic-polar residues
- Modeller: compare.top script

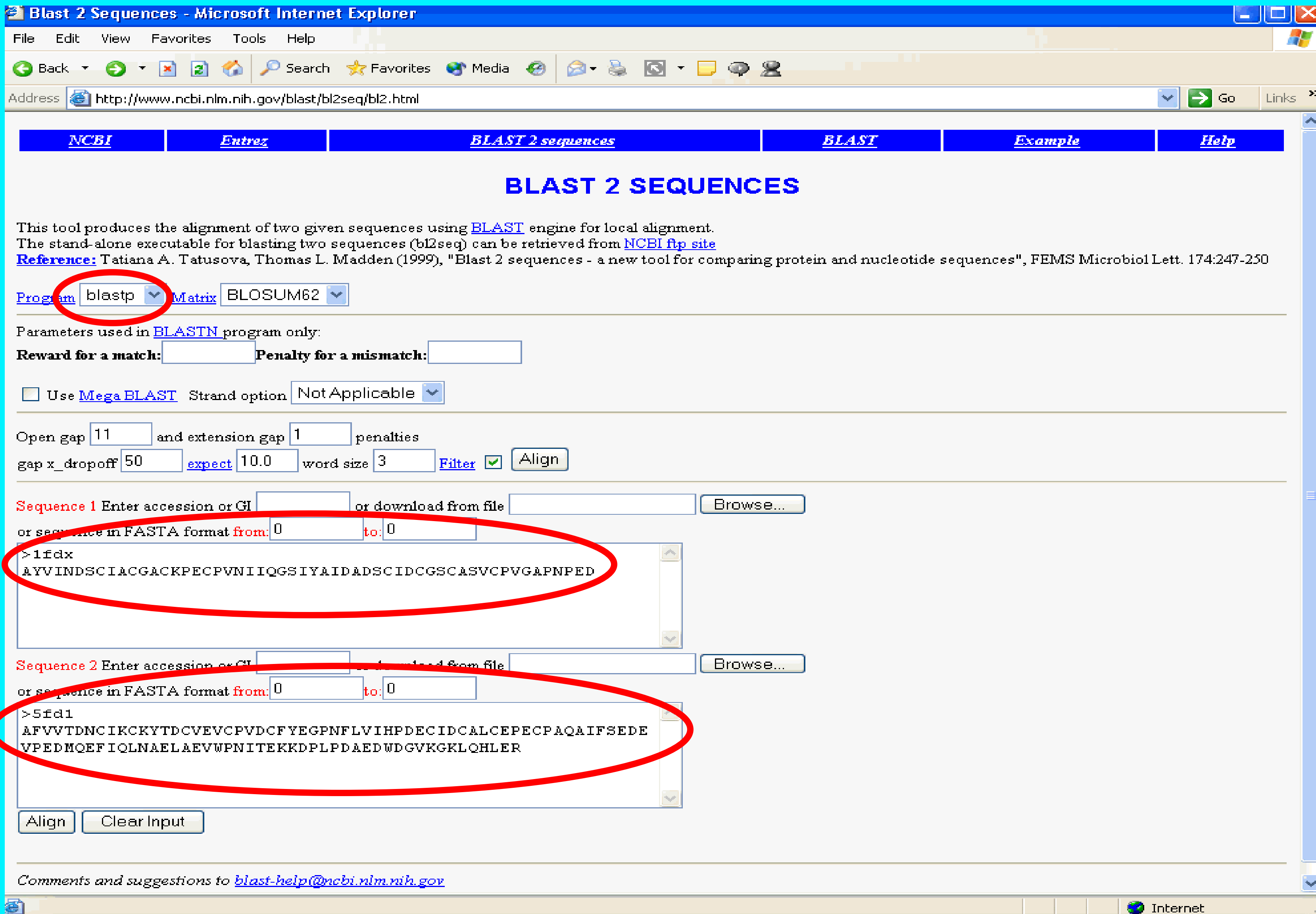
Generating the alignment file

Sequence that we want to model: 1fdx

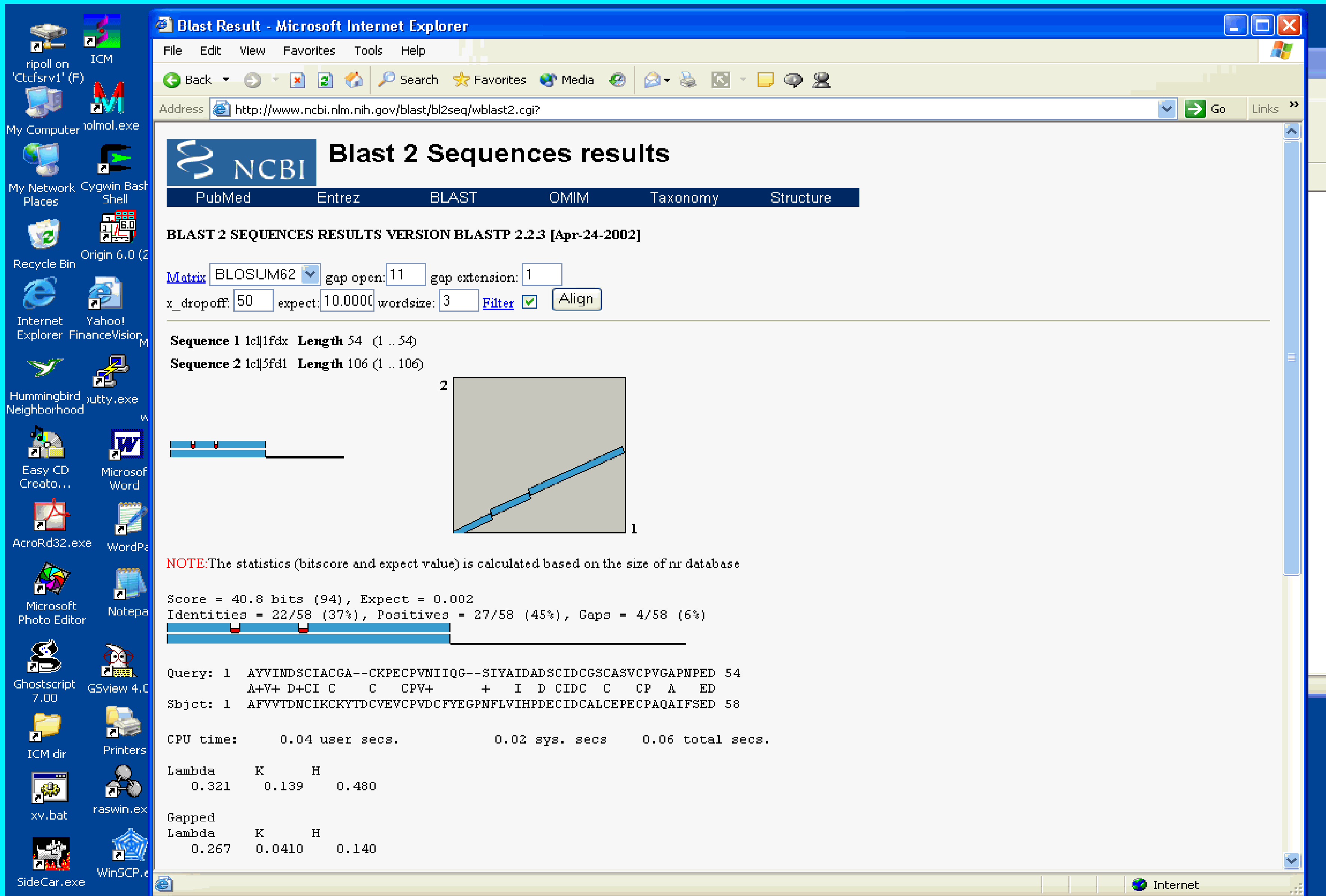
- AYVINDSCIACGACKPECPVNIIQGSIYAIDADS CIDCGSCASVCPVGAPNPED

Sequence of the homologous structure 5fd1

- AFVVTDNCIKCKYTDCVEVCPVDCFYEGPNFL VIHPDECIDCALCEPECPAQAIFSEDEVPEDMQE FIQLNAELAEVWPNITEKKDPLPDAEDWDGVK GKLQHLER



Generate alignment (BLAST the two sequences)



Generating an alignment:
result of BLASTing the two sequences

From BLAST

Query: 1 **AYVINDSCIACGA--CKPECPVNIIQG--SIYAIDADSCIDCGSCASVCPVGAPNPED** 54

Sbjct: 1 **AFVVTDNCIKCKYTDCVEVCPVDCFYEGPNFLVIHPDECIDCALCEPECPAQAIFSED** 58

To PIR

C; A sample alignment in the PIR format; used in tutorial
>P1;5fd1
structureX:5fd1: 1 : :106 : :ferredoxin:Azotobacter vinelandii:1.90:
0.19
AFVVTDNCIKCKYTDCVEVCPVDCFYEGPNFLVIHPDECIDCALCEPECPAQAIFSEDEVPEDMQEFIQLNAELA
EVWPNITEKKDPLPDAEDWDGVKGKLQHLER*
>P1;1fdx
sequence:1fdx :1 : :54 : :ferredoxin:Peptococcus aerogenes: 2.00:-1.00
AYVINDSC--IACGACKPECPVNIIQGS--IYAIDADSCIDCGSCASVCPVGAPNPED-----------------
-------------------------------*

Write a PIR file using the alignment from BLAST