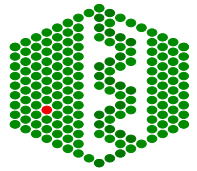
WORKSHOPS



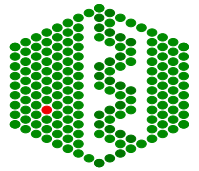
Protein sequence analysis workshop

EMBOSS Package:

Available via www at hgmp or EBI or
www.uk/embnet.org/Software/EMBOSS

Protein seq analysis programs:

Antigenic	Pepcoil
Digest	Helixturnhelix
IEP	Prophecy
Pepinfo	Profit
Pepstats	Prophet
Sigcleave	Tmap



Building a profile

Get sequences and align them:

% emma

input RPOS_* (or seqret them into a file first)

Build profile from alignment using prophecy:

% prophecy

input x.aln file

choose [F]

Use matrix to search SW with profit *

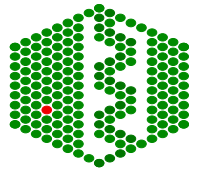
% profit

matrix name from above

input sw:* (or eg sw:*_human)

Retrieve matches, add results to seq file, align, remake profile and rerun till convergence

* Can use same parameters used to create profile, or defaults



Other profiles

Building a Gribskov profile

File x.aln from before

% prophecy

choose [G]

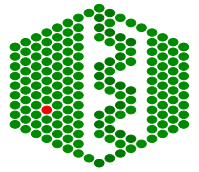
Use matrix to search SW with prophet

% prophet

matrix name from above

input sw:*

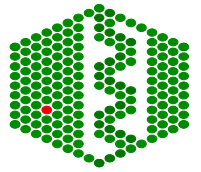
Compare the two different matrices and results of searching



Other input and search options

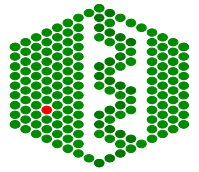
- Input own file with sequences one after the other
- Have list file of sequence names, create fasta file- eg seq.list with sw:opsd_annoc, sw:opsd_apine etc. make fasta file: seqret @seq.list –outseq <outfile>
- Input sequences direct from db with sw:opsd_* or sw:opsd_a* * -any character string, ? -any character

- Can search subset of SW with sw:*_human
- Can search a file of sequences eg. Put together a file of GPCRs



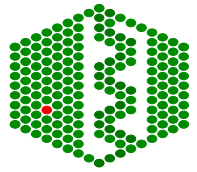
Protein properties analysis

- Run antigenic using A85A_MYCTU.txt
- Run charge using any sequence
- Run digest using ACC8_HUMAN
- Run IEP using any sequence
- Run pepinfo
- Run pepstats



Protein sequence features

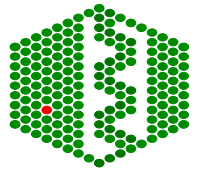
- Run helixturnhelix using LACI_ECOLI.txt
- Run pepcoil using ACC8_HUMAN
- Run tmap using ACC8_HUMAN or gpcr2_aln.txt
- Run sigcleave using signal_asg.txt



Web-based protein analysis tools

- Expasy Proteomics tools http://www.expasy.org.tools
- PredictProtein http://embl-heidelberg.de/predictprotein/

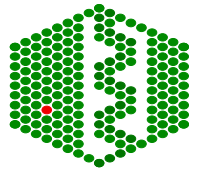
- Use different sequences in directory to analyse, including glycosylation sites etc



Protein sequence analysis workshop

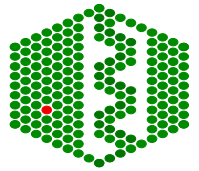
GCG Package:

motifs uses the PROSITE database to find patterns in protein sequences.
profilescan uses a database of profiles to find structural motifs in proteins.
peptidesort shows peptides from a digest of an amino acid sequence.
isoelectric plots the charge as a function of pH for any peptide sequence.
peptidemap creates peptide map of an amino acid sequence.
pepplot makes parallel plot of protein 2ry structure and hydrophobicity.
peptidestructure predicts 2ry structure for a peptide, used by 'plotstructure'.
plotstructure plot output of 'peptidestructure'.
moment makes contour plot of helical hydrophobic moment of a peptide sequence.
helicalwheel plots a peptide structure as a helical wheel.

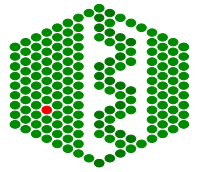


Building a profile with GCG

- Build profile using profilemake and SW:MCM5_*
- Use this to search using profilesearch
- Make alignment of new sequences using profilesegments



Take a sequence and find out as much as possible about its features using different tools



Protein pattern database workshop

PROGRAMS:

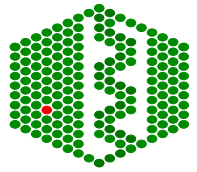
EMBOSS- Patmat, Pfscan

InterProScan

BLOCKS

CDD

Web: Member databases (SMART)



Blocks analysis

- Done via web http://blocks.fhcrc.org/blocks
- Or by email: blocks@blocks.fhcrc.org
- Paste sequence (end4_myctu) into composer, can add comments with #

Searching options:

– Database to search:
  - #DB PLUS(default) | MINUS(PLUS without biased blocks) | PRINTS

– Query sequence type:
  - #TY AUTO(default) | AA | DNA

– For DNA queries, strands to search:
  - #ST BOTH(default) | FORWARD | REVERSE or 2 | 1 | -1

– For DNA queries, genetic code to use for translation:
  - #GE 0(default) to 8

Post-processing options:

– Output type:
  - #OU ALL(default) | SUM | GFF | OLD | RAW

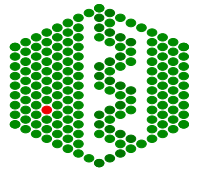
– Output format:
  - #FO TEXT(default) | HTML

– Expected value cutoff:
  - #EX n (default=5)

Sequence definition

#SQ

sequence in fasta or other common formats



EMBOSS

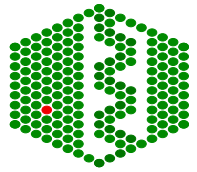
- Pattern matching in Prosite

  % patmatmotifs –full

  Input sw:5NTD_HUMAN
- Finding Fingerprints

  % pfscan

  Input sw:5NTD_HUMAN



InterProScan

Run the individual sequences END4_MYCTU.txt and END4_MYCLE.txt

./InterProScan.pl <seqfile> + ipr

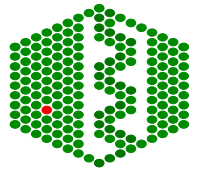
cd tmp/xx

gmake raw –j1 –k

(4 different formats)

gmake txt (xml, html)

Look at different results files or formats



InterProScan cont.

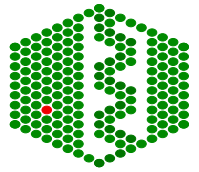
Compare M.tb and M.lep results with diff (txt)

diff file1 file2 (need to specify directory)

Try run diff on raw files

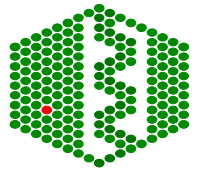
Improve with ./FS_diff.pl <file1> <file2> (if in same directory)

If time permits run Mtb5prot.txt –5 sequences in a file



CDD

- Web server: http://www.ncbi.nlm.nih.gov/Structure/cdd
- Paste sequence in and search (end4_myctu) compare results to InterProScan, search CDD by keyword for related sequences



WEB SEARCHES

Send sequences to InterProScan (http://www.ebi.ac.uk/interpro/scan.html) and member databases

Prosite http://www.expasy.ch/prosite

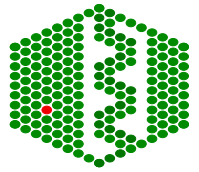
Prints http://www.bioninf.man.ac.uk/dbbrowser/PRINTS/

Pfam http://www.sanger.ac.uk/Software/Pfam/index.shtml

SMART http://smart.embl-heidelberg.de/

ProDom http://www.toulouse.inra.fr/prodom.html

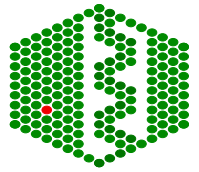
Browse additional features of databases



Complete annotation of proteins

- Take hypothetical proteins from M. tuberculosis:
  - SW- mychyp_seq.txt
  - TRnew- mychyp_trseq.txt

Annotate as completely as possible. For SW compare with the SW annotation (mychyp_sw.txt)



Building Rules

- Collect related protein sequences eg from an InterPro entry into a file (same DR lines)
- Write script to write and count occurrence of DE, CC, KW and FT lines
- Try to find lines common to all entries, build a rule for new sequences hitting the same pattern databases