

# ***Analyzing Protein Sequences***



Modified(!) by Dietlind Gerloff  
for use in BME110/BIOL181  
Winter 2009

© Wiley Publishing, 2007. All Rights Reserved.

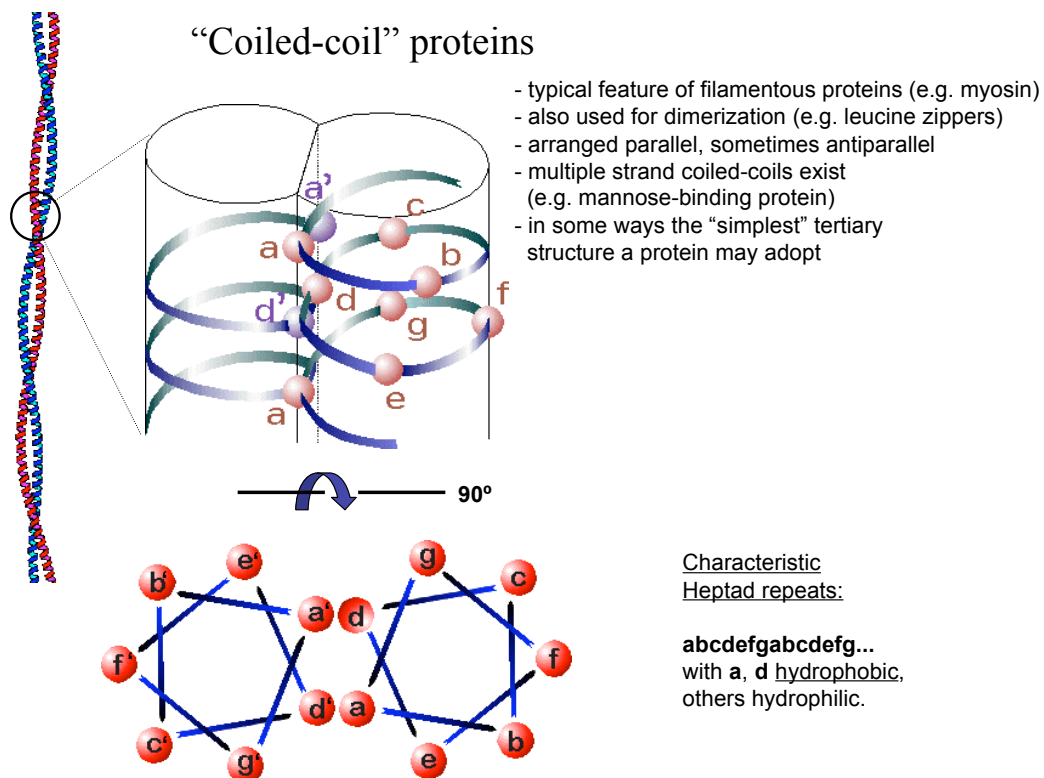


## ***In-silico Biochemistry***

- ✓ Online servers exist to determine many properties of your protein sequences
  - Molecular weight
  - Extinction coefficients
  - Half-life
- ✓ It is also possible to simulate protease digestion
- ✓ All these analysis programs are available on
  - [www.expasy.ch](http://www.expasy.ch)

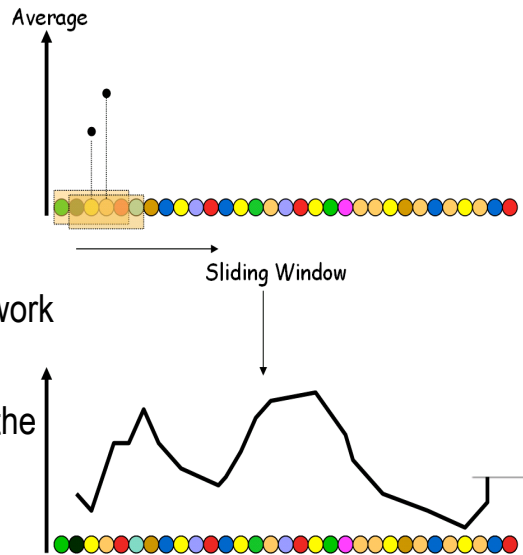
## Analyzing Local Properties

- ✓ Many local properties are important for the function of your protein
  - Hydrophobic regions are potential transmembrane domains
  - Coiled-coiled regions are potential protein-interaction domains
  - Hydrophilic stretches are potential loops
  
- ✓ You can discover these regions
  - Using sliding-window techniques (easy)
  - Using prediction methods such as hidden Markov Models (more sophisticated)



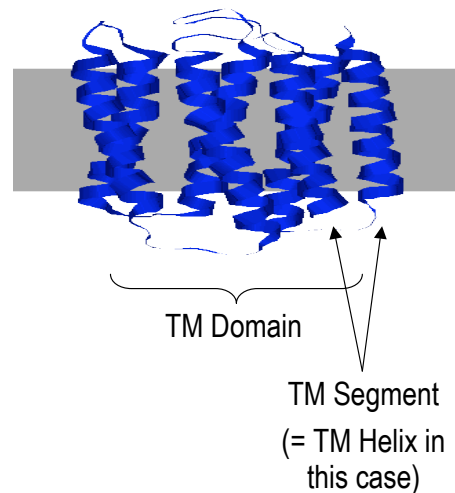
## Sliding-window Techniques

- ✓ Ideal for identifying strong signals
- ✓ Very simple methods
  - Few artifacts
  - Not very sensitive
- ✓ The way most sequence programs work
- ✓ Make the window the same size as the feature you're looking for



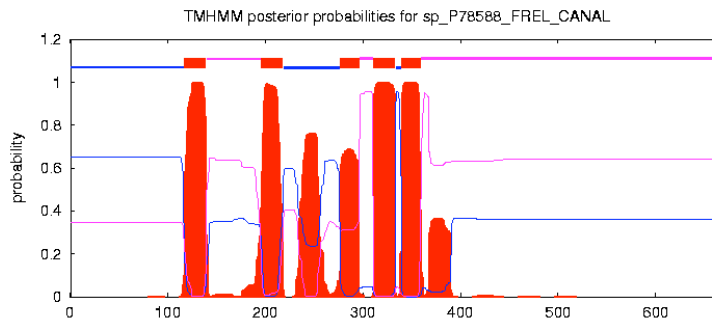
## Transmembrane Segments

- ✓ Discovering a transmembrane domain tells you a lot about your protein
- ✓ Many important receptors have seven transmembrane segments (forming one "domain")
- ✓ There are various prediction programs offering to help detect TM regions
- ✓ The most accurate predictions (at least for bacteria) come from using TMHMM



## Using TMHMM

- ✓ TMHMM is the best current method for predicting transmembrane domains - it works best for bacterial sequences
- ✓ TMHMM uses a “Hidden Markov Model” (HMM)
- ✓ TMHMM output is a prediction (i.e. less trustworthy than an annotation)

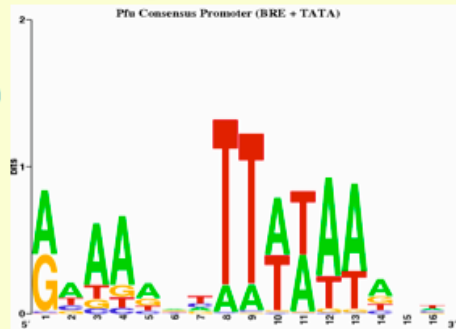


## Predicting Post-translational Modifications

- ✓ Post-translational modifications often occur on similar motifs in different proteins
- ✓ PROSITE is a database containing a list of known motifs, each associated with a function or a post-translational modification
- ✓ Searches of this sort are called “pattern matching” - in PROSITE there are motifs that are highly diagnostic (“low-frequency motifs”) and others that are rather trying to help you pin-point possible sites of, for example, glycosilation (“high-frequency” motifs). The latter can be turned on, or off, in your search.
- ✓ You can search PROSITE by looking for each motif it contains in your protein (the server does that for you!)

# Pattern Matching

- Common task, usually to find short motifs, either
  - protein
    - e.g. key functional site like a catalytic domain
  - nucleic acid
    - e.g. regulatory binding site, like promotor, splice site, etc.



Consensus for *Pyrococcus furiosus* TATA promotor:

[AG]NAANNNTT[TA]{4,4}



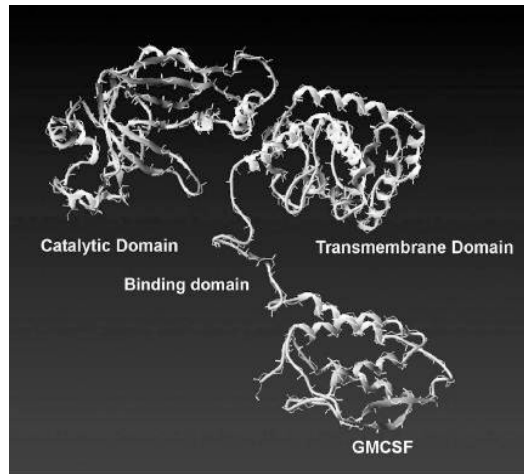
## *Searching for PROSITE Patterns*

- ✓ Search your protein against PROSITE on ExPASy
  - [www.expasy.org/tools/scanprosite](http://www.expasy.org/tools/scanprosite)
- ✓ PROSITE motifs are written as patterns
  - Short patterns are not very informative by themselves
  - They only indicate a possibility (see two slides ago...)
  - Combine them with other information to draw a conclusion
- ✓ Remember also: Not everything is in PROSITE !



## Protein Domains

- ✓ Proteins are usually made of domains
- ✓ A domain is an autonomous folding unit
- ✓ Domains are usually more than 50 amino acids long
- ✓ Different domains may have different functions, e.g.
  - A regulatory domain
  - A binding domain
  - A catalytic domain



Crystal structure of DTGM.

## Discovering Domains

- ✓ Researchers discover domains by
  - Comparing proteins that have similar functions
  - Aligning those proteins
  - Identifying conserved segments
- ✓ A profile is derived from a multiple-sequence alignment (i.e. captures conservation/variation)
- ✓ For each column, the profile indicates which amino acid is more likely to occur

## Domain Collections

- ✓ Scientists have been discovering and characterizing protein domains for more than 20 years
- ✓ 8 collections of domains have been established
  - Manual collections are very precise but small
  - Automatic collections are very extensive but less informative
- ✓ These collections
  - Overlap
  - Have been assembled by different scientists
  - Have different strengths and weaknesses
- ✓ We recommend using several!

## The Magnificent 8

<b>Name</b>	<b>Web Address</b>	<b>Size</b>	<b>Generation</b>
<u>PROSITE-Profile (IP)</u>	<a href="http://www.expasy.org/prosite">www.expasy.org/prosite</a>	616	Manual
<u>PfamA (IP)</u>	<a href="http://www.sanger.ac.uk/Software/Pfam">www.sanger.ac.uk/Software/Pfam</a>	7973	Manual
PRINTs (IP)	<a href="http://www.bioinf.man.ac.uk/dbbrowsers/PRINTS">www.bioinf.man.ac.uk/dbbrowsers/PRINTS</a>	1900	Manual
PRODOM (IP)	<a href="http://protein.toulouse.inra.fr/prodom/current/html/home.php">protein.toulouse.inra.fr/prodom/current/html/home.php</a>	736000	Automatic
<u>SMART (IP)</u>	<a href="http://smart.embl-heidelberg.de">smart.embl-heidelberg.de</a>	685	Manual
COGs	<a href="http://www.ncbi.nlm.nih.gov/COG/new/">www.ncbi.nlm.nih.gov/COG/new/</a>	4852	Manual
TIGRFAM (IP)	<a href="http://www.tigr.org/TIGRFAMs">www.tigr.org/TIGRFAMs</a>	2453	Manual
BLOCKS	<a href="http://blocks.fhcrc.org/">blocks.fhcrc.org/</a>	12542	Automatic

- ✓ Pfam, SMART, and PROSITE-Profiles are most commonly used in protein bioinformatics (don't take the size values here all too seriously but it is true that Pfam is the largest collection amongst the three)

## Searching Domain Collections

- ✓ Pfam-identified domains often include proteins with known functions
- ✓ A match between your protein and a Pfam-entry is desirable
  - A match is a potential indication of a function
  - This is **VERY** informative for further research!
- ✓ Three servers exist to compare proteins and domain collections:
  - InterProScan [www.ebi.ac.uk/interproscan](http://www.ebi.ac.uk/interproscan)
  - CD-Search [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)
  - Motif Scan [www.ch.embnet.org](http://www.ch.embnet.org)

Note: every time the slide says 'domain collections' it really means 'profile/HMM collections'. It is not universally true that one MSA-based profile/HMM corresponds to one domain in a protein structure - it can be less than that, or more.

VEY

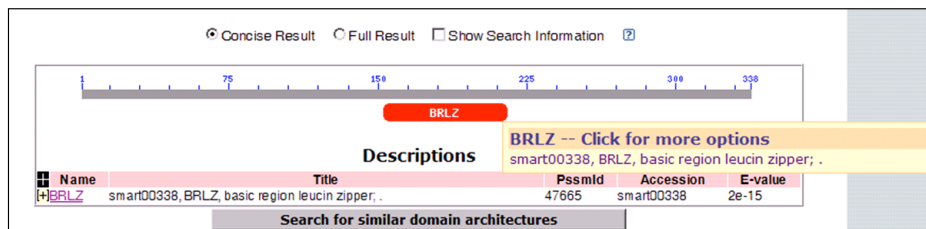
## Using InterProScan

- ✓ InterProScan is the most comprehensive search engine for domain databases
- ✓ Makes it possible to compare alternative results on most collections
- ✓ Does not provide a statistical score

SEQUENCE: FOSB_HUMAN CRC64: DDF827C5047850F LENGTH: 338 aa		
InterPro IPR00209 Domain	Peptidase S8 and S53, subtilisin, kexin, sedolisin PS00138  SUBTILASE_SER	
InterPro IPR00837 Family	Fos transforming protein PR00042  LEUZIPRFOS	
InterPro IPR004827 Domain	Basic-leucine zipper (bZIP) transcription factor SM00338  BRLZ PS00036  BZIP_BASIC PS50217  BZIP	
InterPro IPR008917 Domain		Eukaryotic transcription factor, DNA-binding SSF47454  Euk_transcr_DNA
InterPro IPR011700 Domain		Basic leucine zipper PF07716  bZIP_2

## The CD-Search Output

- ✓ CD search is less extensive than that of InterProScan
- ✓ CD search provides access to the COGs
- ✓ Results come with a statistical evaluation (E-value)
  - $10^{-15}$       **Low E-value**      Good match
  - 2.1            **High E-value**      Bad match



## Predicting Functions with Profiles

- ✓ Finding a match with a Pfam-entry in which a particular catalytic function occurs is good news . . . but what, exactly, does it mean?
- ✓ A match indicates that your sequence has the domain structure . . . but does it also have the function?
- ✓ You cannot say before looking into these details:
  - Where are the catalytic residues on the domain?
  - Does your sequence have the right residues at these positions?