

# **Protein Domains: Prediction from Amino Acid Sequence**

**I690/B680: Structural Bioinformatics**

February 15, 2006

# Introduction

- protein domain identification, given only a sequence, is an important problem
  - gap between sequence and structure data is increasing
- Usual steps given a new sequence
  - if homology can be identified by sequence alignment (to domain databases) the problem is typically easy
  - if homology cannot be reliably identified, the problem is much more difficult
- Databases to look at: Pfam, SMART etc.

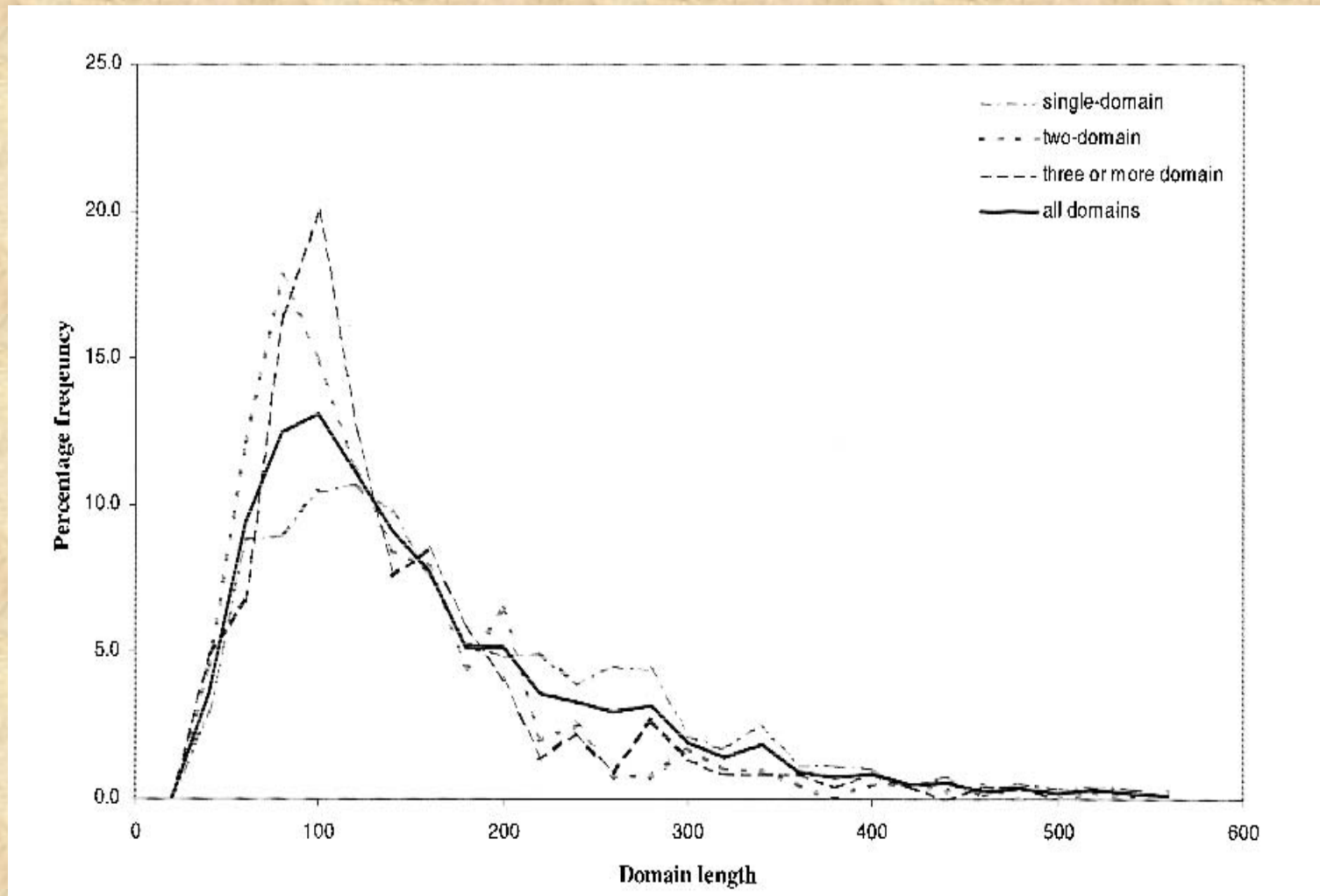
# Practical Importance

- cost-reduction of experimental methods
  - X-ray crystallography
  - NMR spectroscopy
- to help local multiple sequence alignment
  - difficulties arise when whole sequences with “shuffled” domains need to be aligned
- to help fold recognition methods
  - it is easier to look at a sequence in terms of its constituent domains than as a whole chain

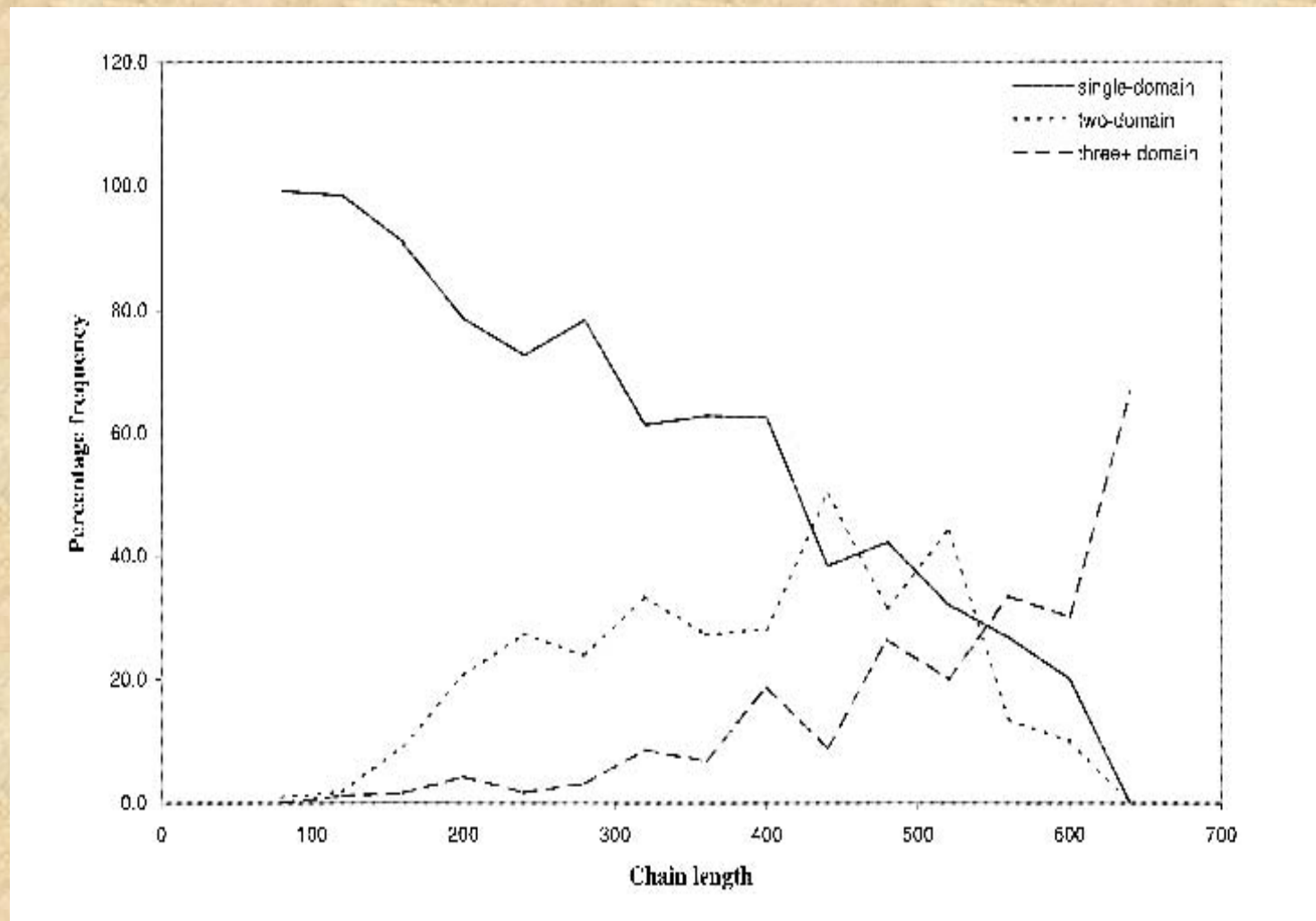
# DomSSEA: idea

- alignment of (predicted) secondary structures can be used for sequence alignment when identity is low
- prediction of secondary structure is sufficiently accurate to be useful (77% or higher; 3-state)
- **Idea:**
  - use Secondary Structure Element Alignment (SSEA)
  - it has been used for fold recognition when no detectable homology existed (using standard sequence alignments)
- **Goal:**
  - to have a fast method capable of predicting on whole genomes
  - result: DomSSEA

# Analysis: Domain Lengths



# Analysis: Domain Lengths



=> Domains can be predicted using the size of proteins

# DomSSEA: algorithm

- predict secondary structure on a new (query) protein
  - PSIPRED (Jones, 1999)
- align query sequence of predicted secondary structures to all sequences (of domains) in the database
  - dynamic programming (McGuffin et al, 2001)
  - adjusted scoring scheme (Przytycka et al, 1999)
- collect top hit from the pair with the highest score
  - use PSI-BLAST to find homologs of the best hit
- make a cut within a coil region

# Accuracy

## Standard measures

- accuracy in domain number prediction
- accuracy of domain boundary prediction
  - $\pm 20$  residues from the CATH assignment  $\Rightarrow$  correct

**Table 1.** *Prediction of one, two, or three or more domain chains*

Prediction of number of domains	% Correctly assigned			
	All	1 domain	2 domains	3 or more domains
PUU	79.0	81.0	66.0	65.0
DomSSEA observed secondary structure	75.4	83.9	47.5	38.1
DomSSEA predicted secondary structure	73.3	82.3	46.0	36.5
DGS-M	76.7	99.8	1.0	0.0
DGS-W	76.7	100.0	0.0	0.0
Absolute difference in length	66.2	78.4	22.3	38.1
Fasta	60.9	74.9	17.3	7.9
Random (weighted)	61.4	75.8	16.8	6.3
Random (basic)	37.9	45.0	16.8	7.9
Sum of squares	62.0	76.0	17.0	7.0

The percentage of chains given a correctly assigned domain number (top prediction), for single, two, and three or more domain chains, as well as for all chains in the representative set.



# Accuracy

**Table 3.** *Prediction of domain boundaries, given a representative set of two domain protein chains ( $\pm 20$  residues)*

Methods	% Correctly assigned boundaries
PUU	81.8
Consensus	52.5
L/(N-1)	49.5
DomSSEA observed secondary structure	49.5
DomSSEA predicted secondary structure	49.0
DGS-M	46.0
Absolute difference in length	44.6
DGS-W	37.1
FASTA	30.0
Random (weighted)	26.8

# Some Other Methods

- **SnapDRAGON**

- George and Heringa, 2002
- *ab initio* protein folding method: DRAGON
  - predicts secondary structure, makes sequence alignment, then folds each chain based on conserved hydrophobicity
  - each folded chain is evaluated for domain boundaries
  - large number of structures are evaluated for domain consistency
- 72% accuracy for # of domains; 51% overall cut prediction

- **CHOPnet**

- Liu and Rost, 2004
- uses aa composition, many predicted features, sequence alignments and combines them using neural networks
- 69% accuracy for # of domains; 50% cut prediction (for 2 domain) chains