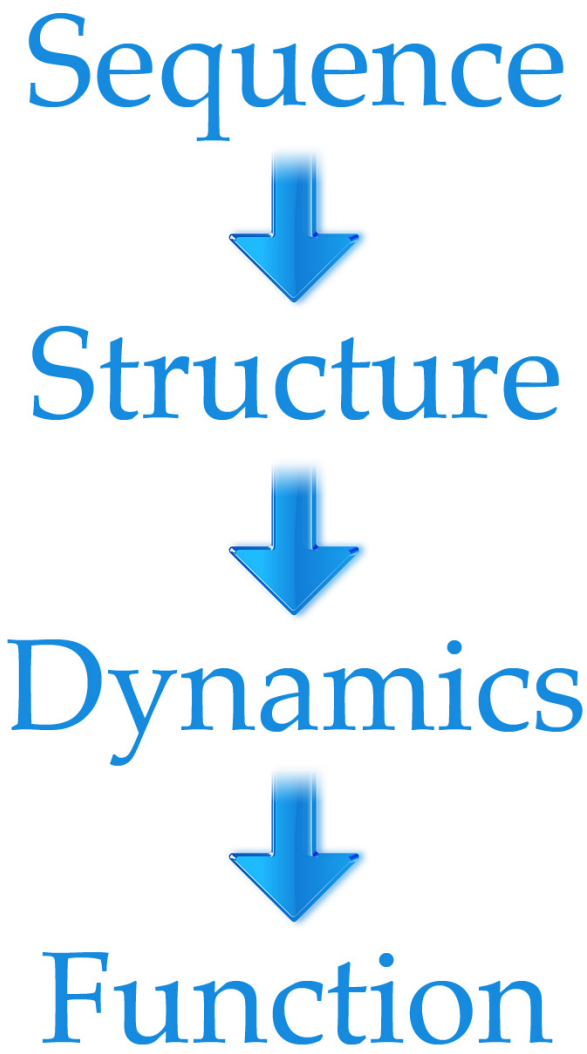
Protein fold classification incorporating additional evolutionary information from phylogenetic profiles

Daniel S. Standage

BCB 569

December 15, 2011

Protein science

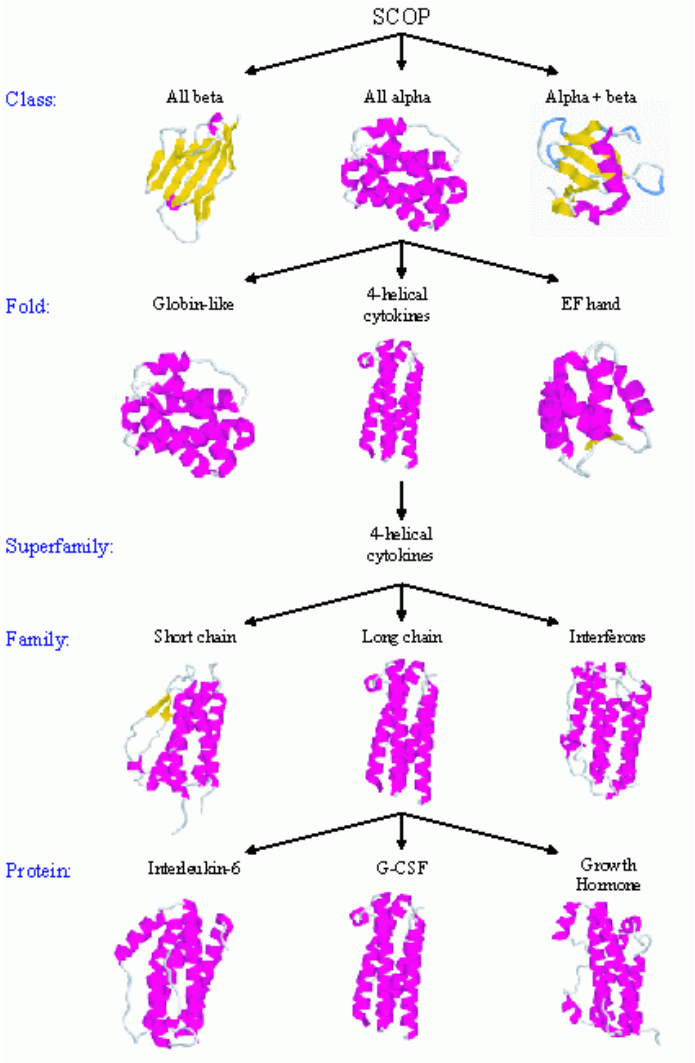


Protein science

Homology modeling

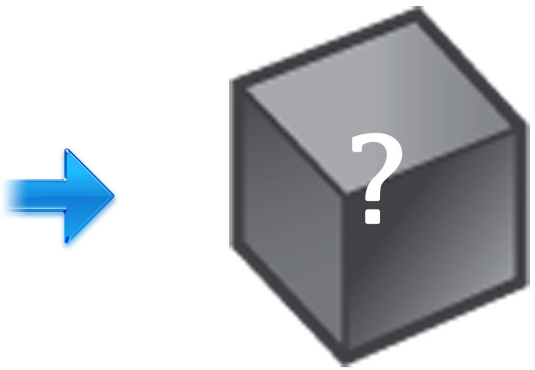
- "sequence $\rightarrow$ structure" problem
- accurate when structure of a close homolog available
- need alternative that does not depend on sequence similarity

Protein fold classification



Protein fold classification

```
>SomeAwesomeProtein
MQPRSERPAGRTQSPEHGSPGPGPEAPPPPPPQPPAPEAERTRPRQARPAAPMEGAVQLL
SREGHSVAHNSKRHYHDAFVAMSRMRQRGLLCDIVLHVAAKEIRAHKVVLASCSPYFHAM
FTNEMSESRQTHVTLHDIDPQALDQLVQFAYTAEIVVGEGNVQTLLPAASLLQLNGVRDA
CCKFLLSQLDPSNCLGIRGFADAHSCSDLLKAAHRYVLQHFVDVAKTEEFMLLPLKQVLE
LVSSDSLNVPSEEEVYRAVLSWVKHDVDARRQHVPRLMKCVRLPLLSRDFLLGHVDAESL
VRHHPDCKDLLIEALKFHLLPEQRGVLGTSRTRPRRCEGAGPVLFAVGGGSLFAIHGDCE
AYDTRTDRWHVVASMSTRRARVGVAAVGNRLYAVGGYDGTSDLATVESYDPVTNTWQPEV
SMGTRRSCLGVAALHGLLYSAGGYDGASCLNSAERYDPLTGTWTSVAAMSTRRRYVRVAT
LDGNLYAVGGYDSSSHLATVEKYEPQVNVWSPVASMLSRRSSAGVAVLEGALYVAGGNDG
TSCLNSVERYSPKAGAWESVAPMNIRRSTHDLVAMDGWLYAVGGNDGSSSLNSIEKYNPR
TNKWVAASCMFTRRSSVGVAVLELLNFPPPSSPTLSVSSTSL
```



Protein fold classification

My project

- implement PFRES method
- extend method with new features

Data

- includes sequences from 27 most populated folds in SCOP
- pairwise sequence similarity $< 35\%$

Data

Training data

- 313 domains (Ding & Dubchak, 2001)

Testing data

- 385 domains (Ding & Dubchak, 2001)
- 908 domains (Shen & Kurgan, 2007)

Features

```
>SomeAwesomeProtein
MQPRSERPAGRTQSPEHGSPGPGPEAPPPPPPQPPAPEAERTRPRQARPAAPMEGAVQLL
SREGHSVAHNSKRHYHDAFVAMSRMRQRGLLCDIVLHVAAKEIRAHKVVLASCSPYFHAM
FTNEMSESRQTHVTLHDIDPQALDQLVQFAYTAEIVVGEGNVQTLLPAASLLQLNGVRDA
CCKFLLSQLDPSNCLGIRGFADAHSCSDLLKAAHRYVLQHFVDVAKTEEFMLLPLKQVLE
LVSSDSLNVPSEEEVYRAVLSWVKHDVDARRQHVPRLMKCVRLPLLSRDFLLGHVDAESL
VRHHPDCKDLLIEALKFHLLPEQRGVLGTSRTRPRRCEGAGPVLFAVGGGSLFAIHGDCE
AYDTRTDRWHVVASMSTRRARVGVAAVGNRLYAVGGYDGTSDLATVESYDPVTNTWQPEV
SMGTRRSCLGVAALHGLLYSAGGYDGASCLNSAERYDPLTGTWTSVAAMSTRRRYVRVAT
LDGNLYAVGGYDSSSHLATVEKYEPQVNVWSPVASMLSRRSSAGVAVLEGALYVAGGNDG
TSCLNSVERYSPKAGAWESVAPMNIRRSTHDLVAMDGWLYAVGGNDGSSSLNSIEKYNPR
TNKWVAASCMFTRRSSVGVAVLELLNFPPPSSPTLSVSSTSL
```

Features

- PCV
- SSC
- nDSSS
- aDSSS
- n
- PP

Profile-based composition vector (PCV)

- search with PSI-BLAST
- generate PSSM a_{ki} (an $L \times 20$ matrix)
- PCV has following form

$$PCV_i = \sum_{k=1}^{L} \frac{\max(a_{ki}, 0)}{L} \qquad (i = 1, 2, ..., 20)$$

Profile-based composition vector (PCV)

	A	R	N	...	Y
1	-3	1	2	...	5
2	-6	5	-6	...	-2
...	...	...	...	...	...
L	7	4	4	...	-3

$$PCV_i = \sum_{k=1}^{L} \frac{\max(a_{ki}, 0)}{L} \qquad (i = 1, 2, ..., 20)$$

Secondary-structure-based features

- three structure categories: H=helix, E=strand, C=coil
- content: **SSC**
- contiguous segments: **DSSS**
- arrangements of contiguous segments: **ADSSS**

Secondary-structure-based features

SSC_m	$m \in \{H, E, C\}$
$DSSS_m$	$m \in \{H, E, C\}$
$ADSSS_m$	$m \in \{H, E, C\}^3$

CCHHHHHHHCCCEEEEECCEEEEEEEEECC

SSC_H = 7 / 30 = 0.233
SSC_E = 14 / 30 = 0.467
SSC_C = 9 / 30 = 0.300

$DSSS_H$ = 1
$DSSS_E$ = 2
$DSSS_C$ = 0

$ADSSS_{HEE}$ = 1

Phylogenetic profile

- captures phylogenetic distribution of protein
- database of n reference genomes $G = \{G_1, G_2, ..., G_n\}$
- phylogenetic profile of a gene g

$$p = p_1 p_2 ... p_n \quad \text{s.t.} \quad p_i = 1 \quad \text{if} \quad g \in G_i$$

Full feature space

Characteristic	# Features
PCV	20
SSC	3
DSSS	3
ADSSS	27
Length	1
Phylogenetic profile	12
Total	**66**

Feature calculation

- PCVs
  - PSI-BLAST
  - NR database
- SS features
  - PSI-PRED
  - Uniref90 database
- phylogenetic profiles
  - new code
  - custom protein databases
- parallel pipeline

Phylogenetic profile

Eukaryotes	**Dicots**
Arabidopsis thaliana	*Arabidopsis thaliana*
Aspergillus nidulans	*Carica papaya*
Cryptosporidium parvum	*Cucumis sativus*
Danio rerio	*Glycine max*
Drosophila melanogaster	*Lotus japonicus*
Eremothecium gossypii	*Manihot esculenta*
Homo sapiens	*Medicago trunculata*
Mus musculus	*Mimulus guttatus*
Physcomitrella patens	*Populus trichocarpa*
Plasmodium falciparum	*Prunus persica*
Saccharomyces cerevisiae	*Ricinus communis*
Zea mays	*Solanum lycoparsicum*

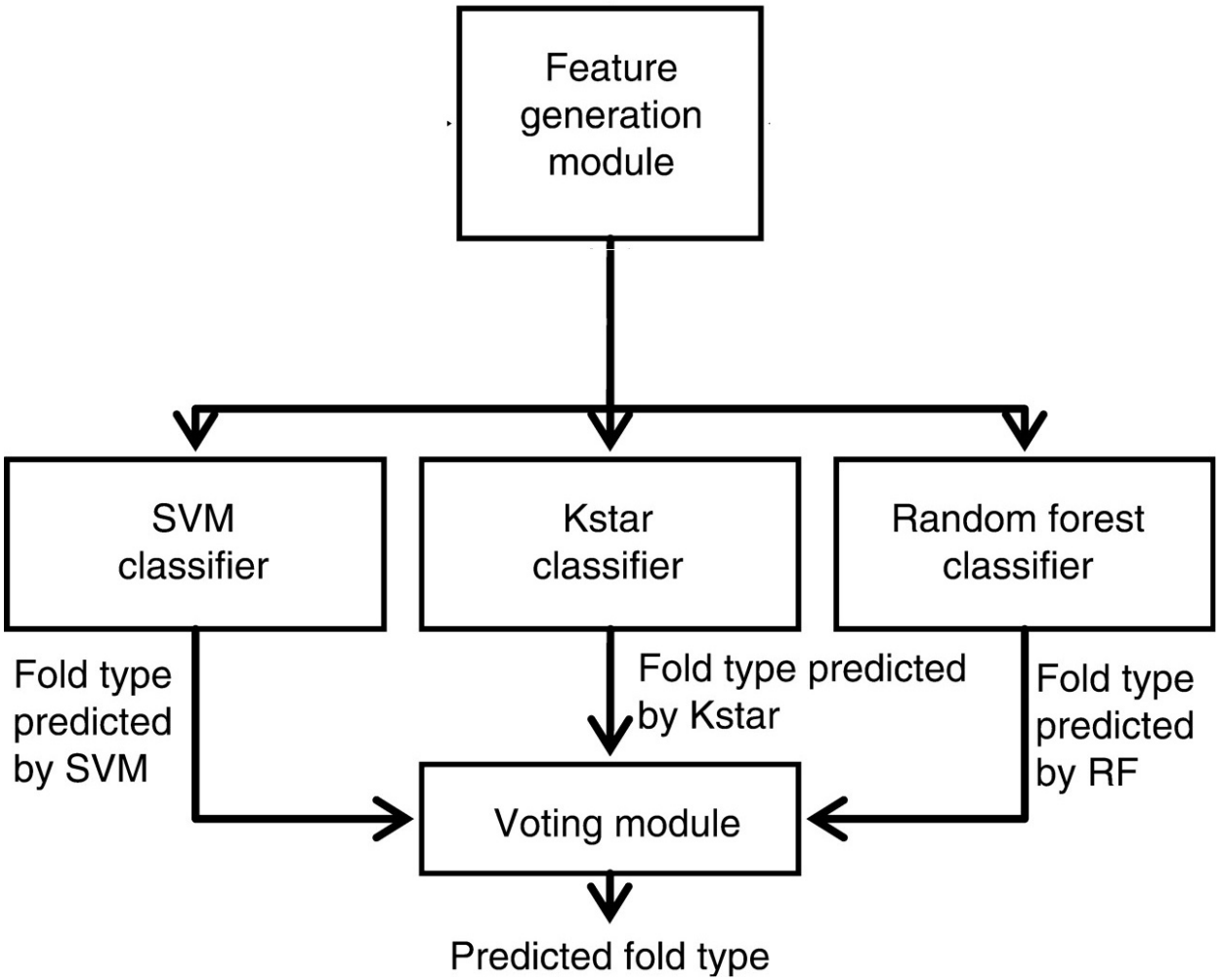
Data set feature spaces

	Training Set	Test Set 1	Test Set 2
PFRES	Tr_P	$T1_P$	$T2_P$
PFRES + Euk.	Tr_{PE}	$T1_{PE}$	$T2_{PE}$
PFRES + Dic.	Tr_{PD}	$T1_{PD}$	$T2_{PD}$

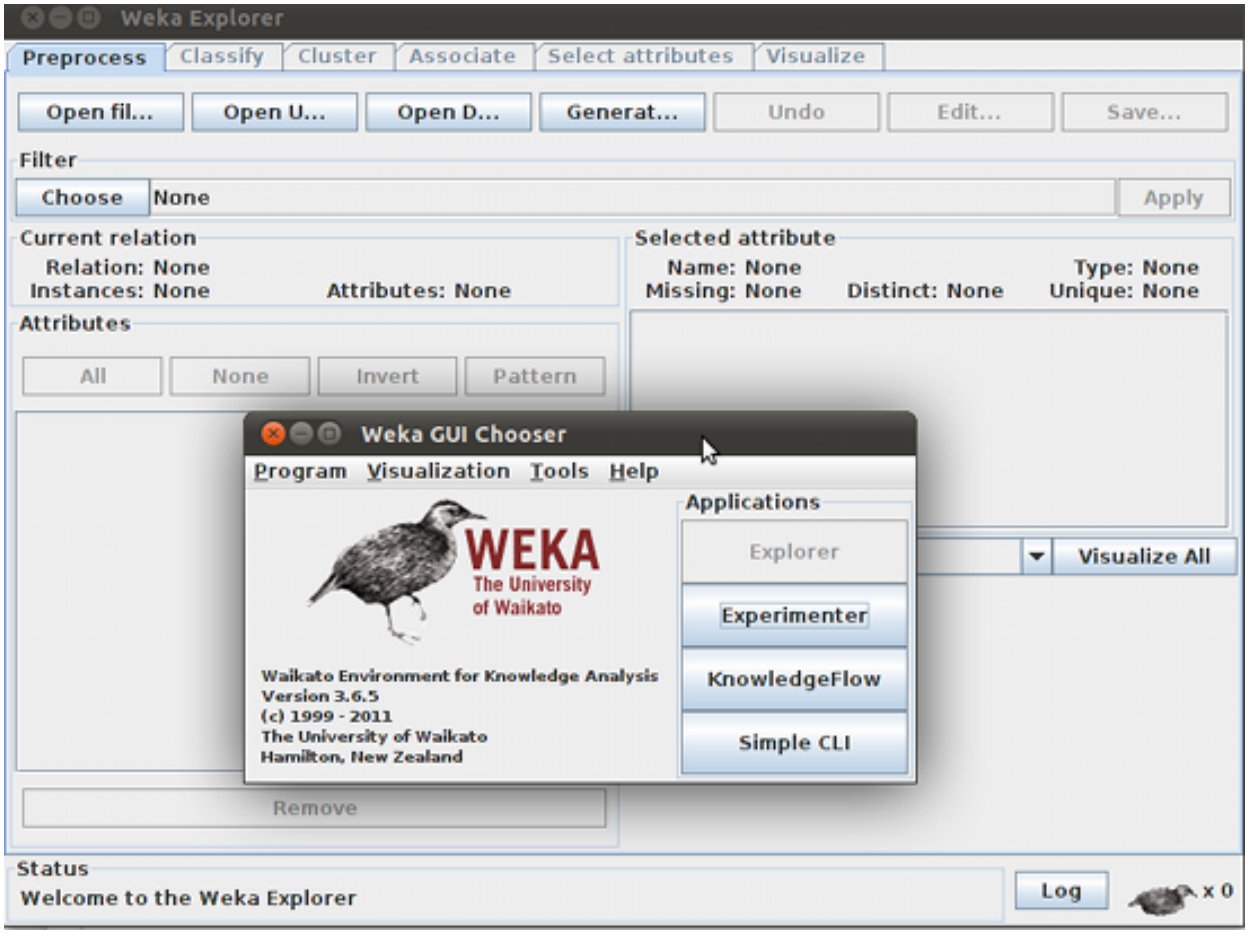
Feature selection: information gain

Characteristic	# Features	Selected Features
PCV	20	20
SSC	3	3
DSSS	3	3
ADSSS	27	10
Length	1	1
Phylogenetic profile	12	6
Total	**66**	**43**

Machine learning



Machine learning



Results

Classifier		Eukaryotes	Dicots	PFRES	Published
RF					
	Test 1	**65.8%**	62.9%	65.5%	66.8%
	Test 2	**64.6%**	59.8%	64.0%	63.3%
SVM					
	Test 1	61.4%	54.8%	66.8%	66.1%
	Test 2	53.1%	51.1%	62.9%	62.4%
Kstar					
	Test 1	**63.2%**	57.2%	63.2%	65.0%
	Test 2	**59.4%**	52.5%	56.7%	62.7%
Ensemble					
	Test 1	65.8%	60.1%	67.1%	68.4%
	Test 2	**62.7%**	56.5%	62.6%	66.4%

Conclusions

- Short phylogenetic profiles do not provide a significant performance improvement.
- My method performed comparably despite the higher-dimensional feature space.
- With more training data and longer phylogenetic profiles, I expect a substantial performance improvement.

Acknowlegements

- Dr. Jernigan
- Dr. Hongbin Shen
- Dr. Lukasz Kurgan

Acknowlegements

Thank you!