Max Planck Institute for Molecular Genetics,
October 21st, 2002

Allergenicity Prediction using Sequence Profiles

M.B. Stadler
University of Bern

Overview

- Allergy
  - What is an allergen/allergic reaction?
- Generalized profiles
  - Modeling sequence motifs
  - Construction of a profile
- Allergenicity prediction
  - Current prediction algorithm (FAO/WHO)
  - Profile-based prediction

Overview

- **Allergy**
  - What is an allergen/allergic reaction?
- **Generalized profiles**
  - Modeling sequence motifs
  - Construction of a profile
- **Allergenicity prediction**
  - Current prediction algorithm (FAO/WHO)
  - Profile-based prediction

Allergens are...

- harmless substances
- (almost) only proteins
- inducing IgE immune response
- rare: ~ 800’000 proteins (Sp/TrEMBL)

↓

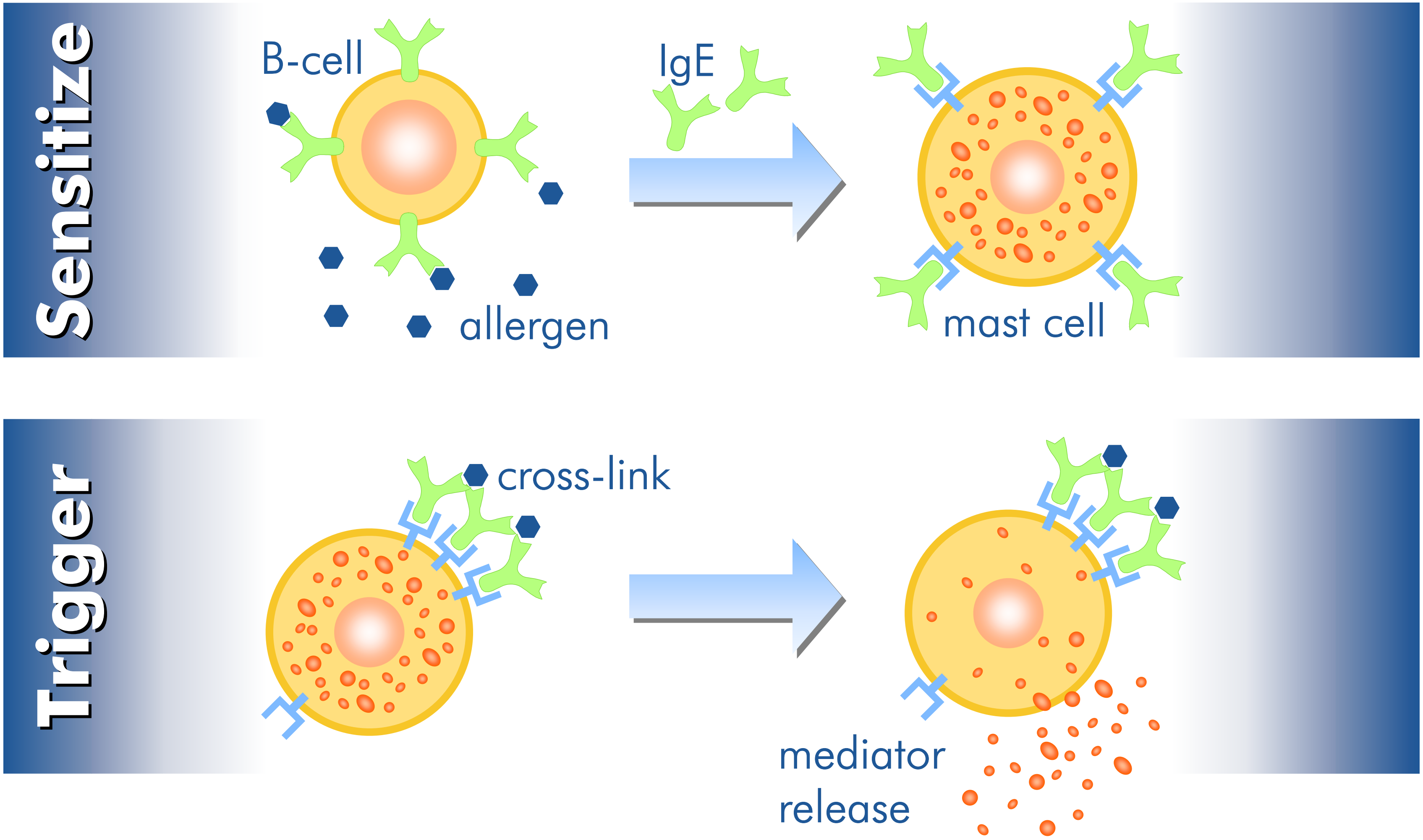
~ 800 sequences
< 100 allergens

1 in 100’000 proteins:

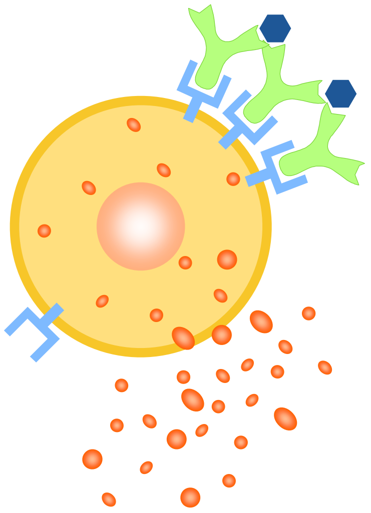
Common Allergens

house dust mites, pollen & spores, pets, insects, milk, eggs, nuts

Immediate Hypersensitivity Reaction



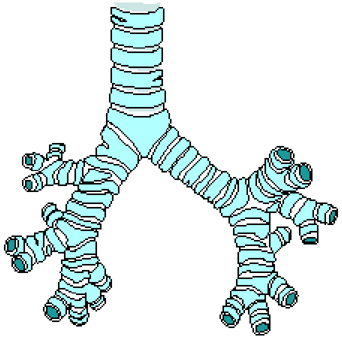
Response to Mediators



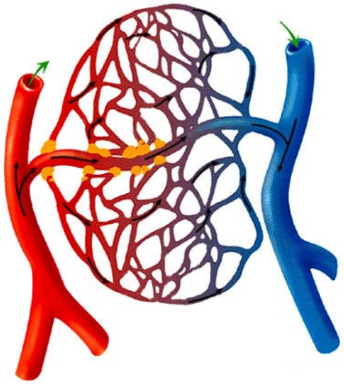
Mediators:

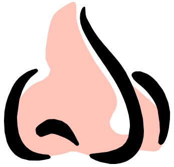
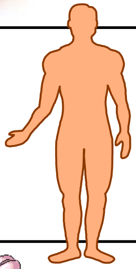
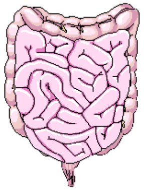
Histamine
Leukotrienes
Cytokines
Chemokines
Enzymes

Bronchi: Constriction



Blood vessels: Vascular leakage (tissue edema)

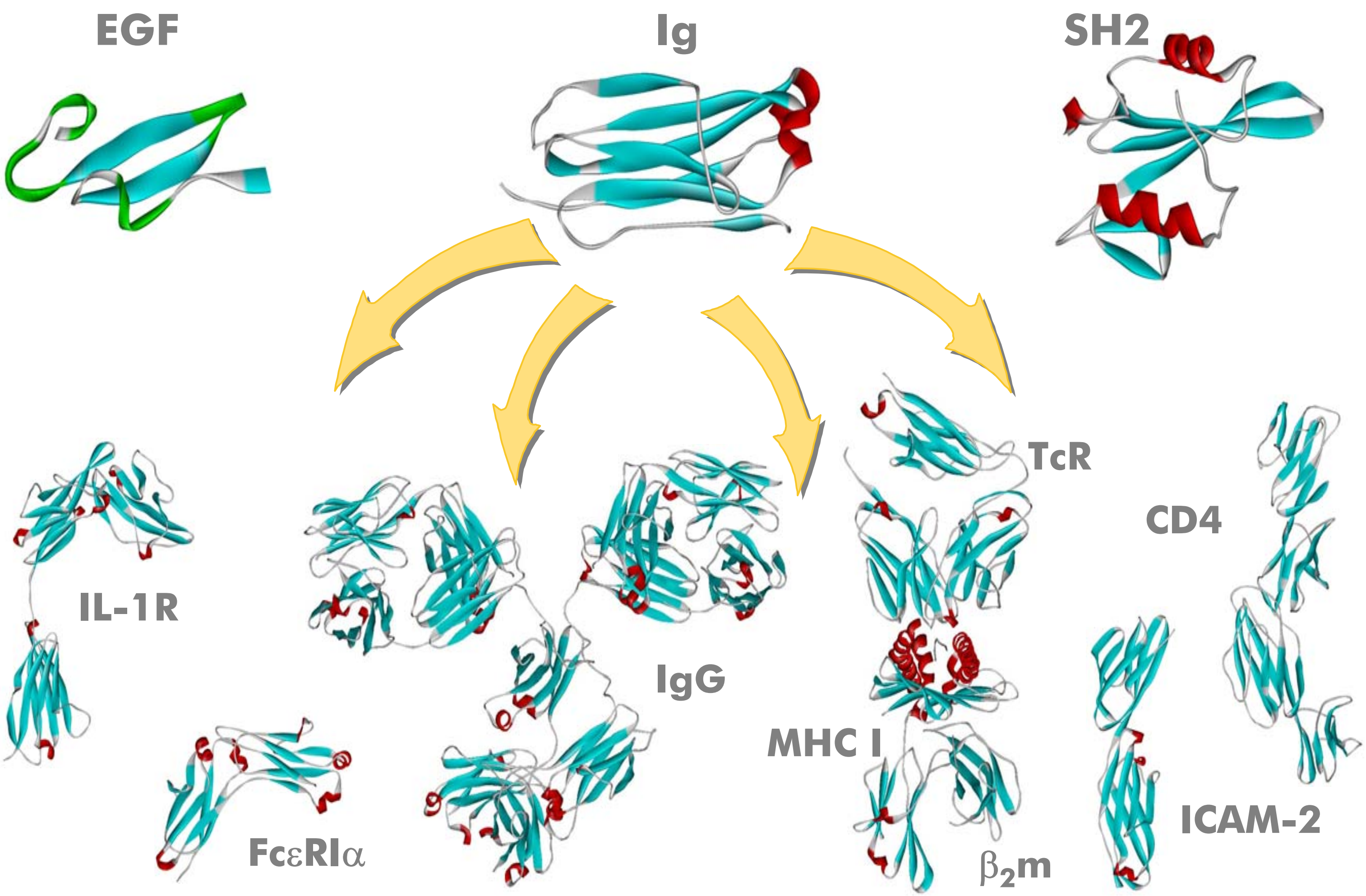


Reaction	Site	Signs/Symptoms
Rhinitis	Nose 	Sneezing, rhinorrhea, nasal itching, congestion
Asthma	Lungs	Coughing, wheezing, shortness of breath
Dermatitis	Skin	Itching, rash
Conjunctivitis	Eye	Itching, redness, tearing
Anaphylaxis	Systemic 	Hypotension, shock, death
Food	Gut 	Bloating, vomiting, diarrhea, cramping

Overview

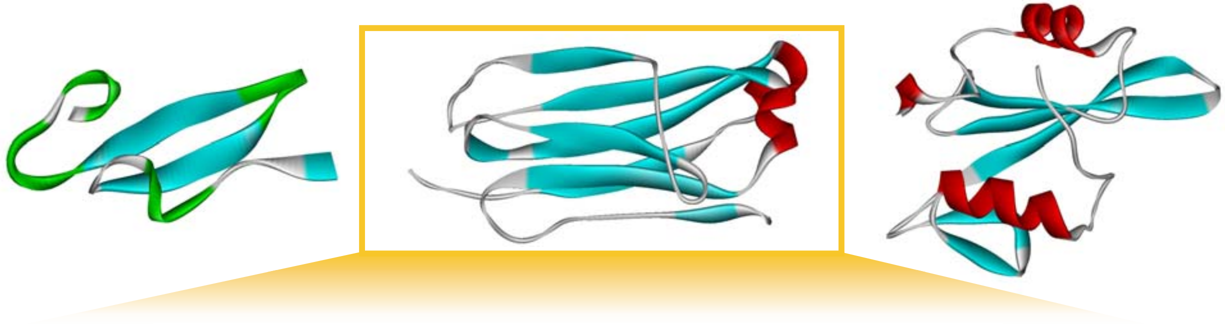
- Allergy
  - What is an allergen/allergic reaction?
- Generalized profiles
  - Modeling sequence motifs
  - Construction of a profile
- Allergenicity prediction
  - Current prediction algorithm (FAO/WHO)
  - Profile-based prediction

Shared Domain Topologies



Sequence-Structure Relationship

Domain (Fold):

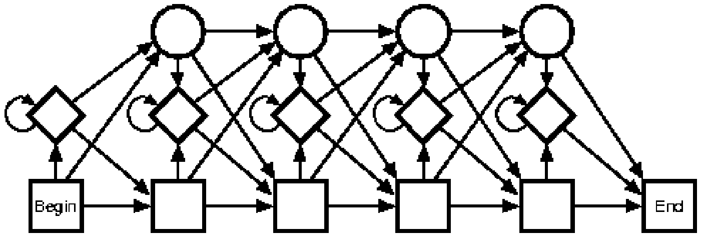


Sequences:

```
- - K R E K K L S L H I A D S Q P - G D S A T -
- - - - G D T F V L T L S D F R R - E N E G Y -
- - - - T L P L T L Q I P Q V S L - Q F A G S -
- - - - - - N G T L K I K H L K T - D D Q D I -
- - - - - - N G D L K I K N L T R - D D S G T -
- D P S W K D G S I V I H N L D Y - S D N G T -
- N M D H K V C N L L L K D L K P - E D S G T -
- - L T Q K A F H L V I S P V R T - E D S A T -
- L W D Q G N F P L I I K N L K I - E D S D T -
```

Sequence Motif:

[FY]-x-C-x-[VA]-x-H



A Generalized Profile

- position-specific match-, gap- and insert-scores
- score distribution over sequence space
- parameters:

begin/end:

B^e B^i
E^e E^i

insertion/match:

I(X) I(*)
M(X) M(*)

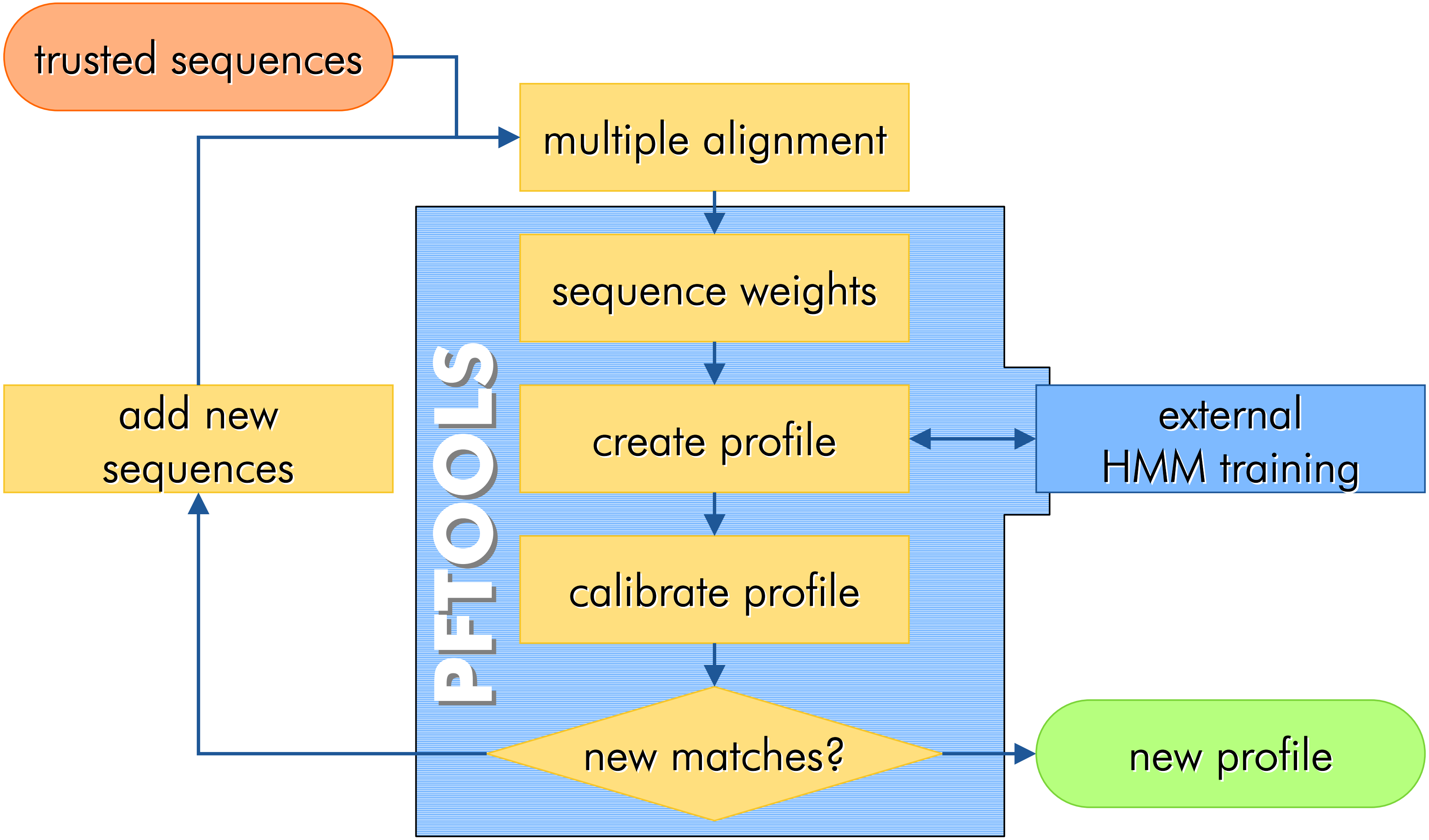
deletion:

D

state transitions:

$T_{B \rightarrow M}$ $T_{B \rightarrow I}$ $T_{B \rightarrow D}$ $T_{B \rightarrow E}$
$T_{M \rightarrow M}$ $T_{M \rightarrow I}$ $T_{M \rightarrow D}$ $T_{M \rightarrow E}$
$T_{I \rightarrow M}$ $T_{I \rightarrow I}$ $T_{I \rightarrow D}$ $T_{I \rightarrow E}$
$T_{D \rightarrow M}$ $T_{D \rightarrow I}$ $T_{D \rightarrow D}$ $T_{D \rightarrow E}$

Manual Construction of Sequence Profiles



trusted sequences

multiple alignment

sequence weights

create profile

external HMM training

calibrate profile

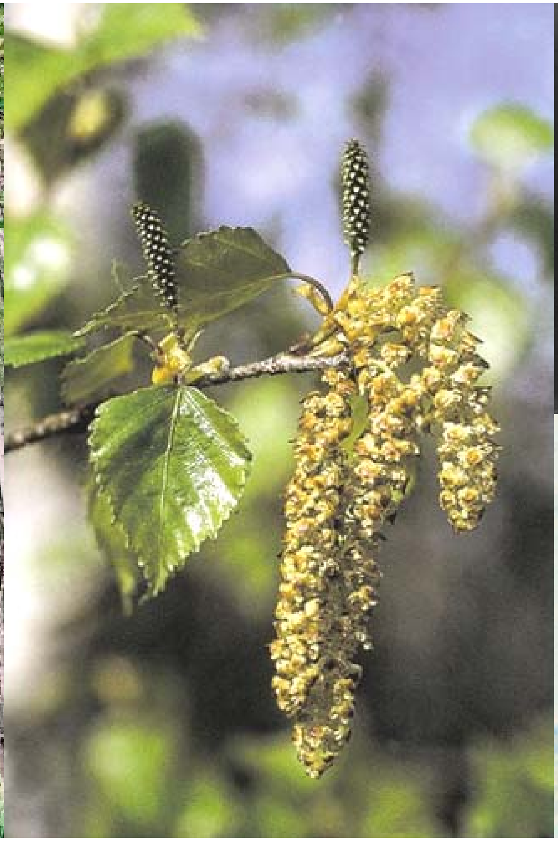
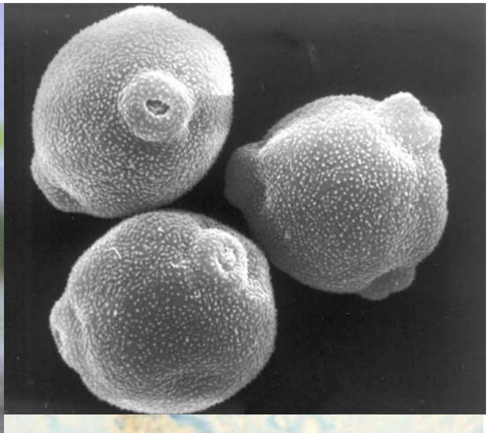
new matches?

new profile

add new sequences

PFTOOLS

The Birch Allergen Family

Clinical cross-reactions include:

birch, celery, carrot, apple, hornbeam, hazelnut, alder, apricot, cherry, pear

Creating a Bet v 1-Profile

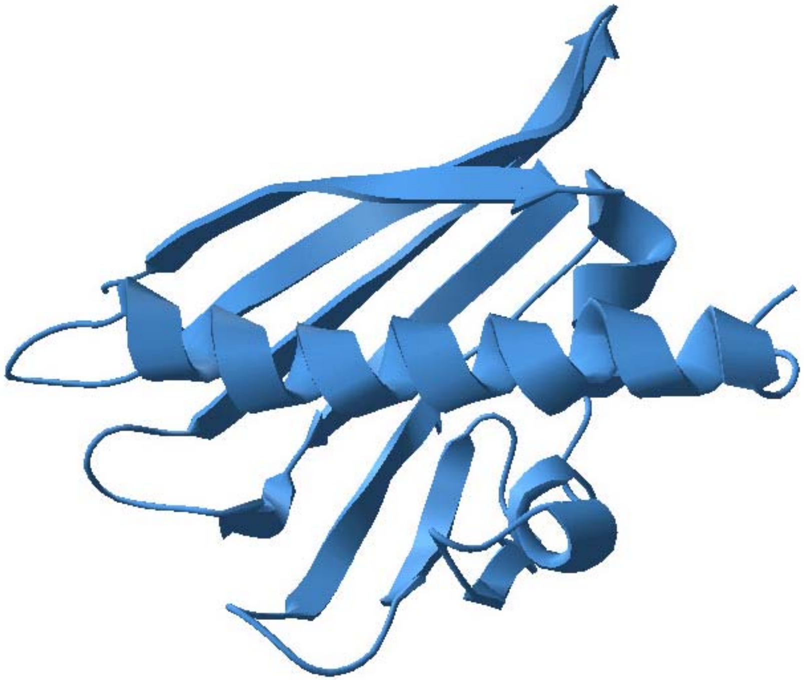
trusted sequences

multiple alignment

extensive literature search:

13 allergens:

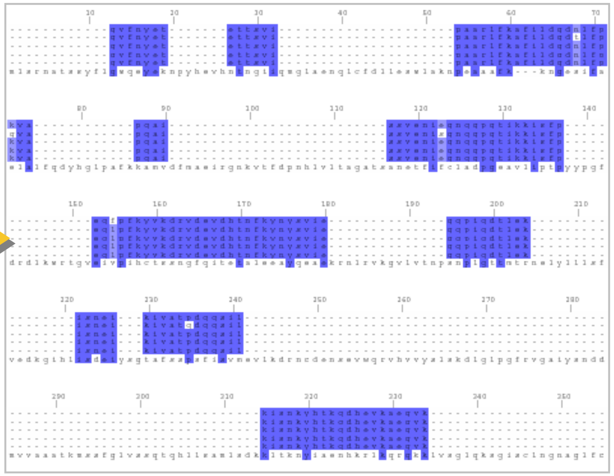
- **6 structures**
- **94 sequences**

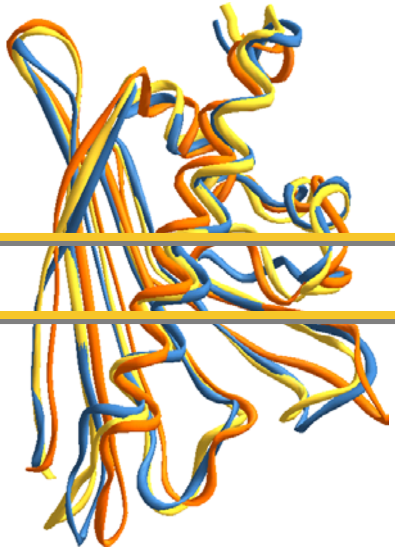


Good and Bad Sequence Alignments

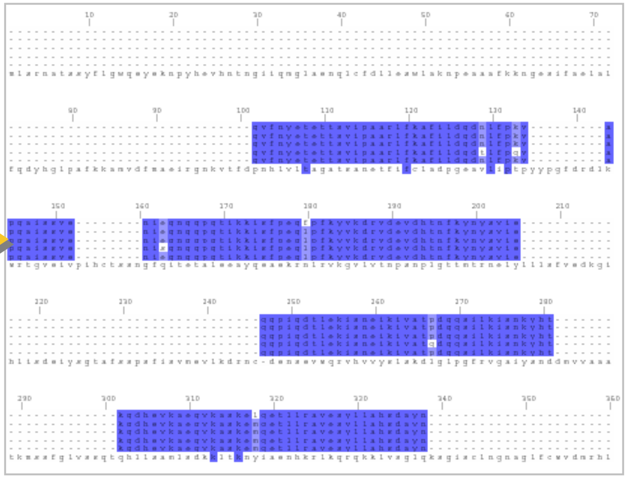
Sequence based alignment

ASTQSPSVFPLTRCCK
NIPSNATSVTLGCLAT
GYFPEPVMVTWDTGSL
NGTTMTLPATTLTLSG
HYATISLLTVSGAWAK
QMFTCRVAHTPSSTDW
VDNKTFSVCSRDFTPP
TVKILQSSCDGGGHFP
PTIQLLCLVSGYTPGT
INITWLEDGQVMDVDL

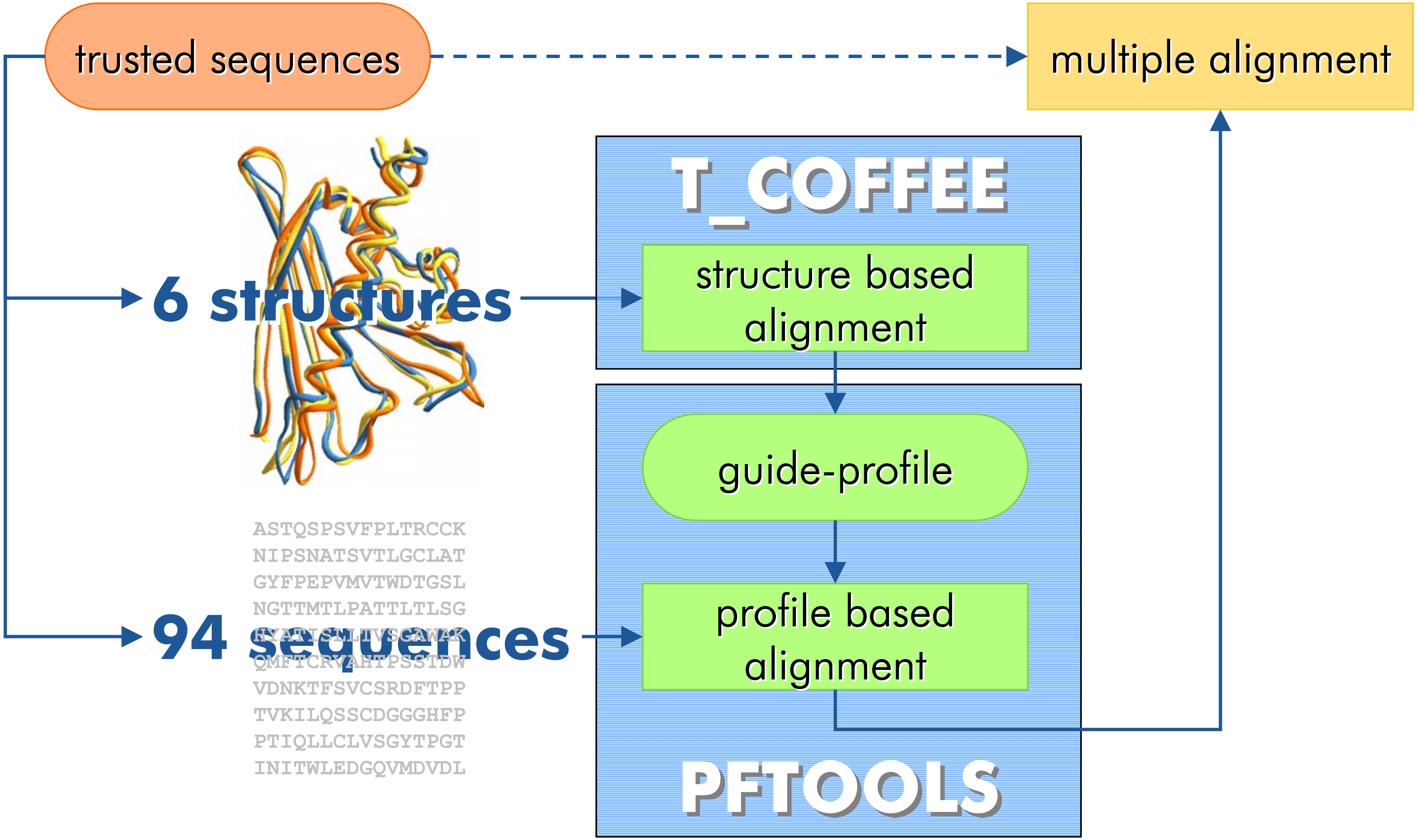




Structure based alignment



Creating a Bet v 1-Profile: Structural Alignment



trusted sequences

multiple alignment

6 structures

94 sequences

ASTQSPSVFPLTRCCK
NIPSNATSVTLGCLAT
GYFPEPVMVTWDTGSL
NGTTMTLPATTLTLSG
HYATISLLTVSGAWAK
QMFTCRVAHTPSSTDW
VDNKTFSVCSRDFTPP
TVKILQSSCDGGGHFP
PTIQLLCLVSGYTPGT
INITWLEDGQVMDVDL

T_COFFEE

structure based alignment

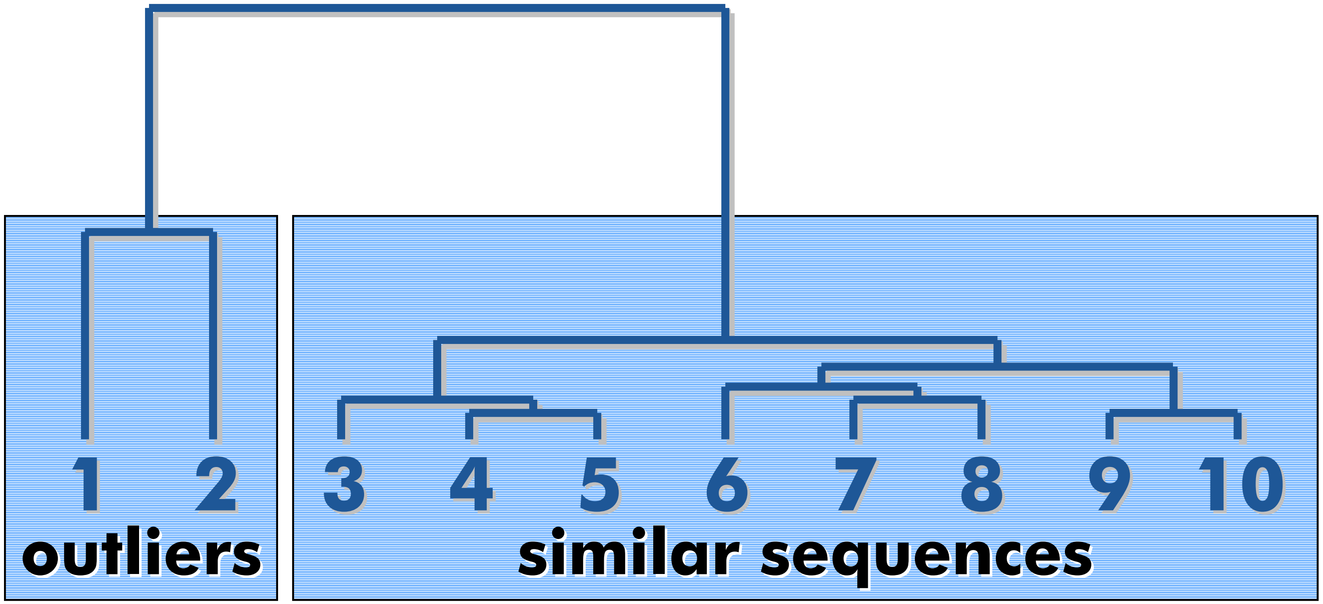
guide-profile

profile based alignment

PFTOOLS

Sequence Weights

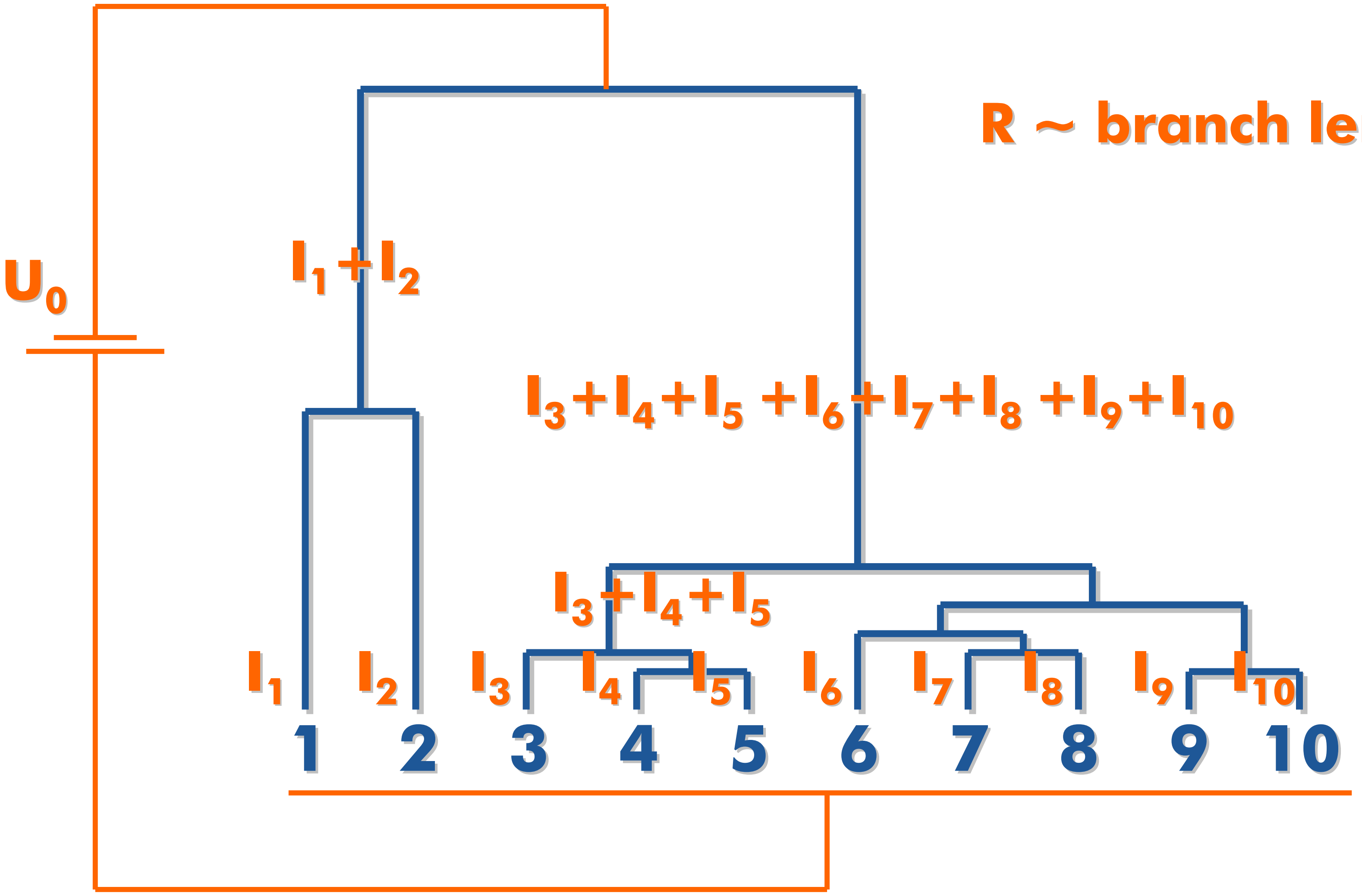
multiple alignment → sequence weights → create profile



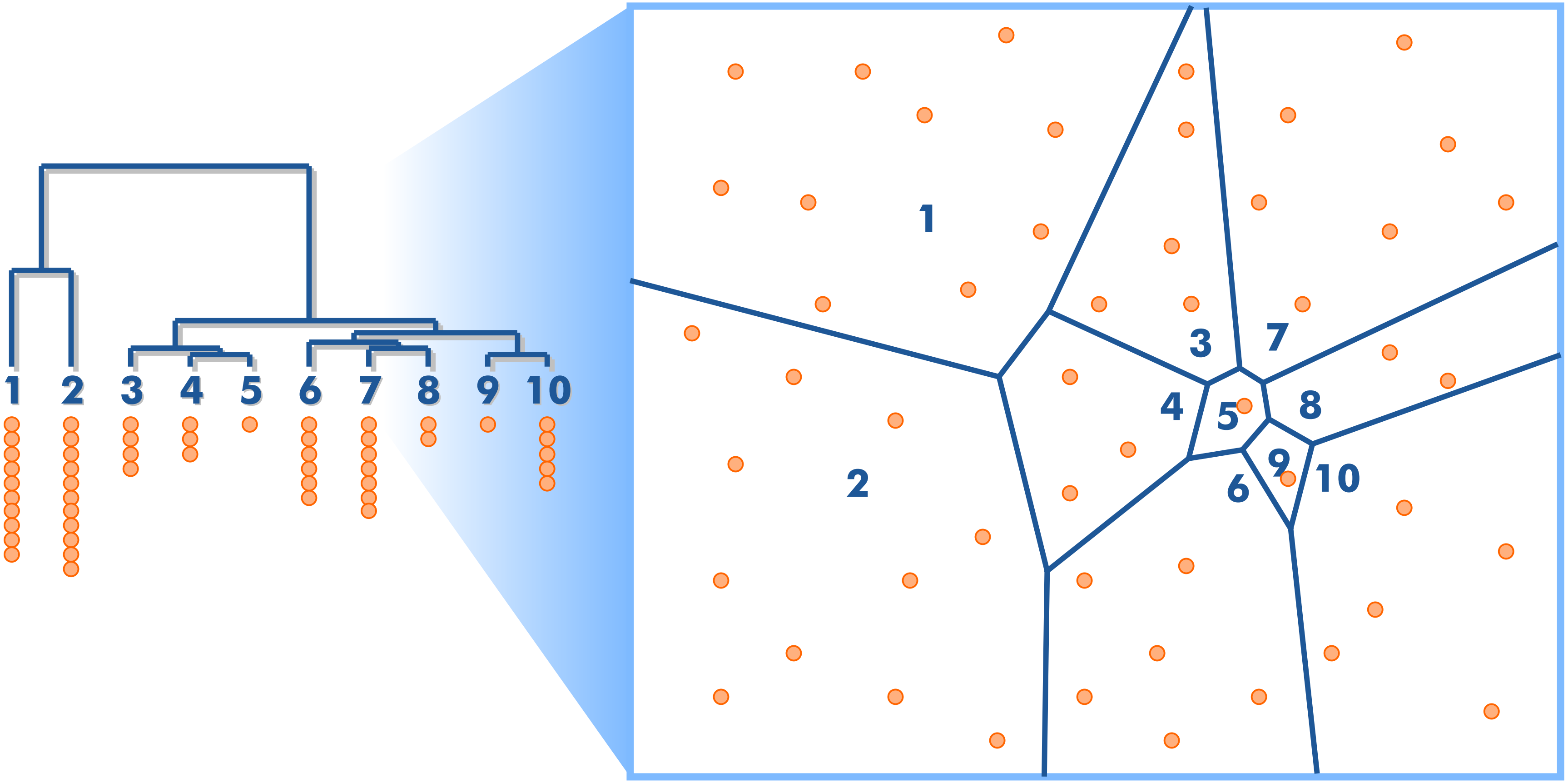
Profile should be:

focused	– –	++
unbiased	++	– –

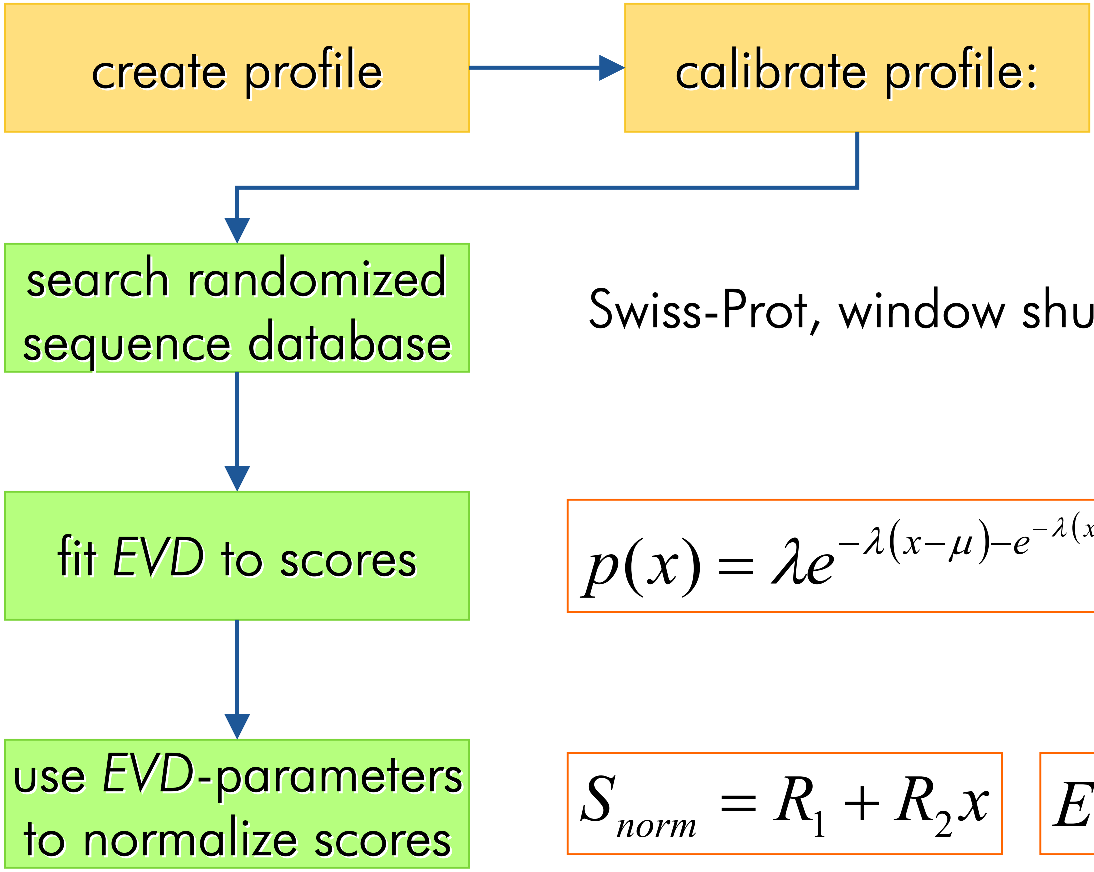
Sequence Weights: Tree Derived



Sequence Weights: 'Voronoi'



Empiric Profile Calibration



create profile

calibrate profile:

search randomized sequence database

fit *EVD* to scores

use *EVD*-parameters to normalize scores

Swiss-Prot, window shuffled

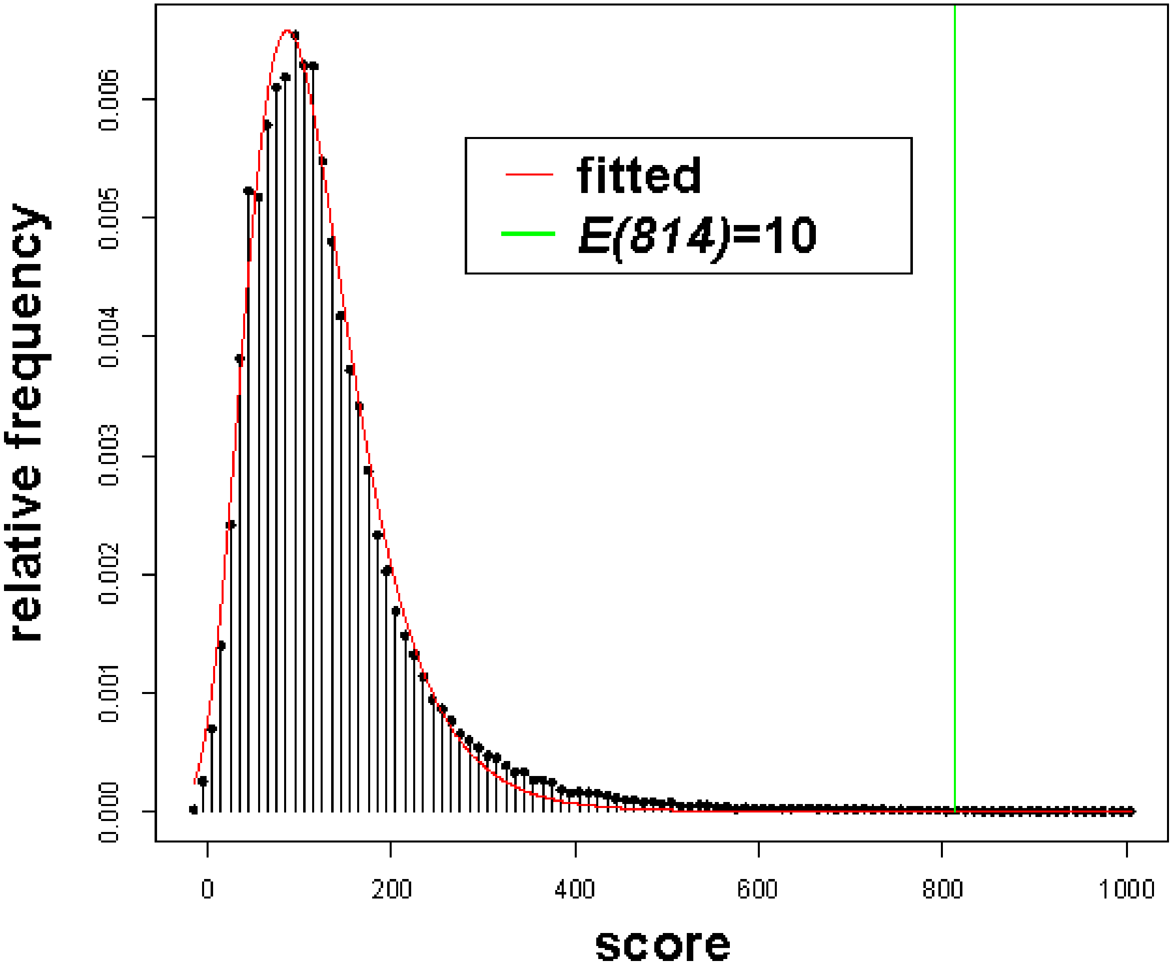
$$\boxed{p(x) = \lambda e^{-\lambda(x-\mu) - e^{-\lambda(x-\mu)}} \cong e^{-\lambda(x-\mu)}}$$

$$\boxed{S_{norm} = R_1 + R_2 x} \quad \boxed{E(x, A) = A \cdot 10^{-S_{norm}}}$$

$$R_1 = \frac{\ln \frac{A}{N} - \lambda\mu}{\ln 10} \qquad R_2 = \frac{\lambda}{\ln 10}$$

x: raw score
A: no. of residues
N: no. of sequences

Normalized Score and E-value



$$E(x, A) = A \cdot 10^{-R_1 - R_2 x}$$

Swiss-Prot 40:

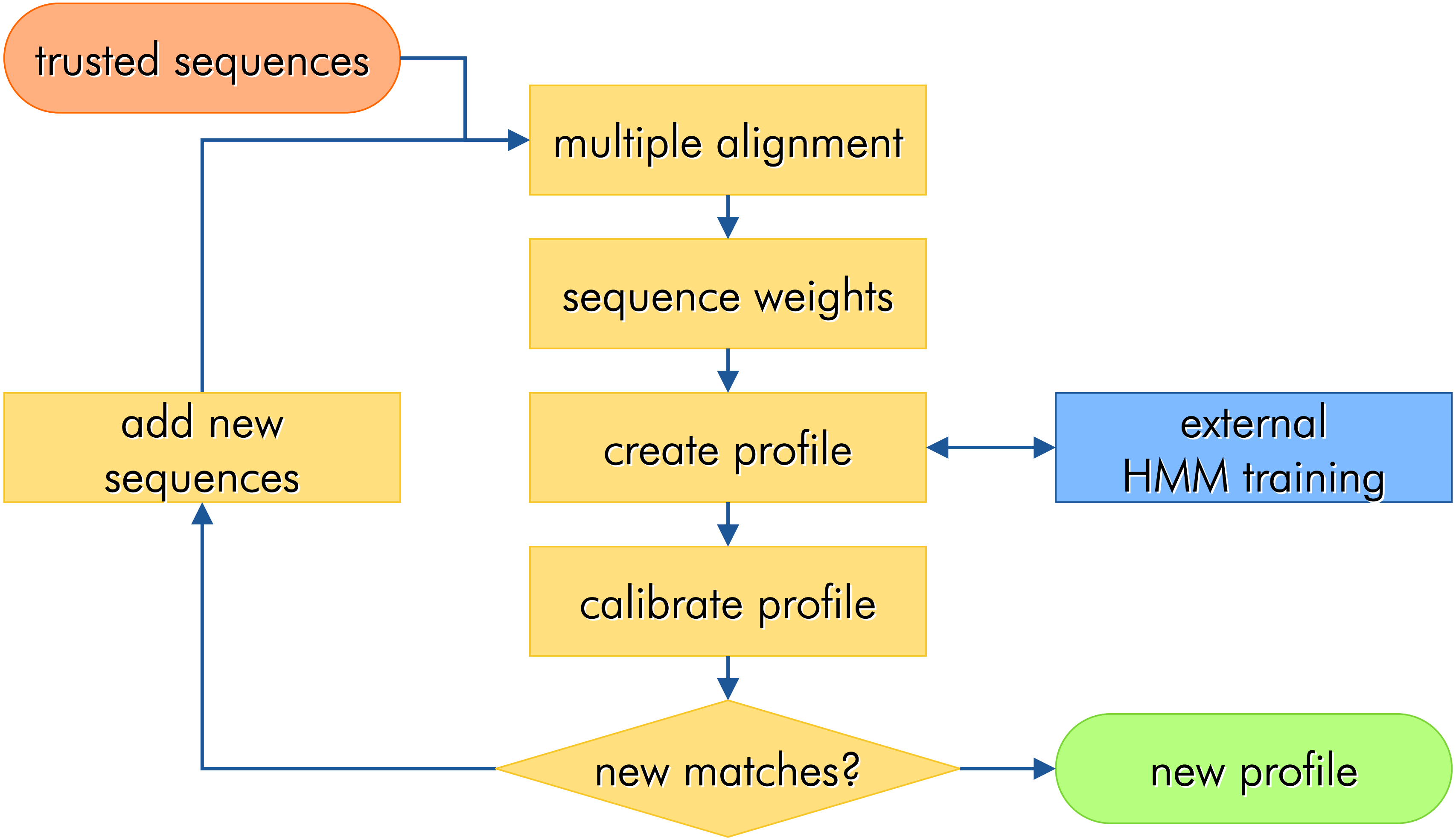
***A*=37'315'215**

***x*=814**

S_{norm}=6.5

E-value=10

Profile Construction is an Iterative Process



FHA Domain Profile: PS50006

Hofmann & Bucher, 1995, TIBS

Database Entry	Profile 1		Profile 2		Profile 3		Profile 4		Profile 5	
	weight	score	weight	score	weight	score	weight	score	weight	score
CDS1_SCHPO		2.3		2.9		2.7		3.9		3.8
FHL1_YEAST		14.5	50.0	40.6	23.5	33.7	14.7	31.9	12.7	23.8
FKH1_YEAST		7.9		19.0	23.9	35.6	13.5	34.1	12.1	24.7
FKH2_YEAST		6.3		18.6	23.2	35.5	13.3	33.8	11.3	24.0
FRAH_ANASP		3.4		5.7		5.1		6.6		7.5
KAPP_ARATH		2.8		5.1		5.3		5.7		5.7
KI67_HUMAN		3.1		2.7		3.6		4.0		4.0
MNF_MOUSE	100.0	49.1	50.0	37.4	29.4	28.2	18.4	27.7	15.6	20.8
SPK1_YEAST		5.0		6.6		6.3		6.9		7.2
YHR5_YEAST		2.5		4.4		5.8		9.0		9.6
YKI5_CAEEL		3.6		6.4		5.1		7.1		7.4
HUMKIAA10_1		3.1		3.1		3.4		3.8		4.8
MLB1770_16		4.8		4.7		5.8		5.1		5.0
SC9346_10		4.1		6.8		4.4		6.3		6.8
SCCXIV38K_16		2.7		3.6		5.5		8.6		9.3
SCD9717_7		4.2		6.5		8.7		11.8	14.3	18.6
SCPPR1GEN_4		3.1		3.1		3.3		2.9		3.2
SPAC17G8_10		6.0		8.5		10.2	20.5	19.2	18.5	17.1
SYCSLRG_6		4.0		5.5		4.8		6.4		6.9
SYCSLRG_63		8.9		6.2		6.5		8.7		9.7
YSCL8083_15		3.7		6.5		5.0		6.4		6.8
YSCL9470_15		8.2		8.9		9.7	19.7	18.9	15.3	20.5
B61188		2.8		3.5		2.9		3.6		3.8
Highest false positive		8.1		8.2		8.7		7.7		7.8

… PS50006 continued

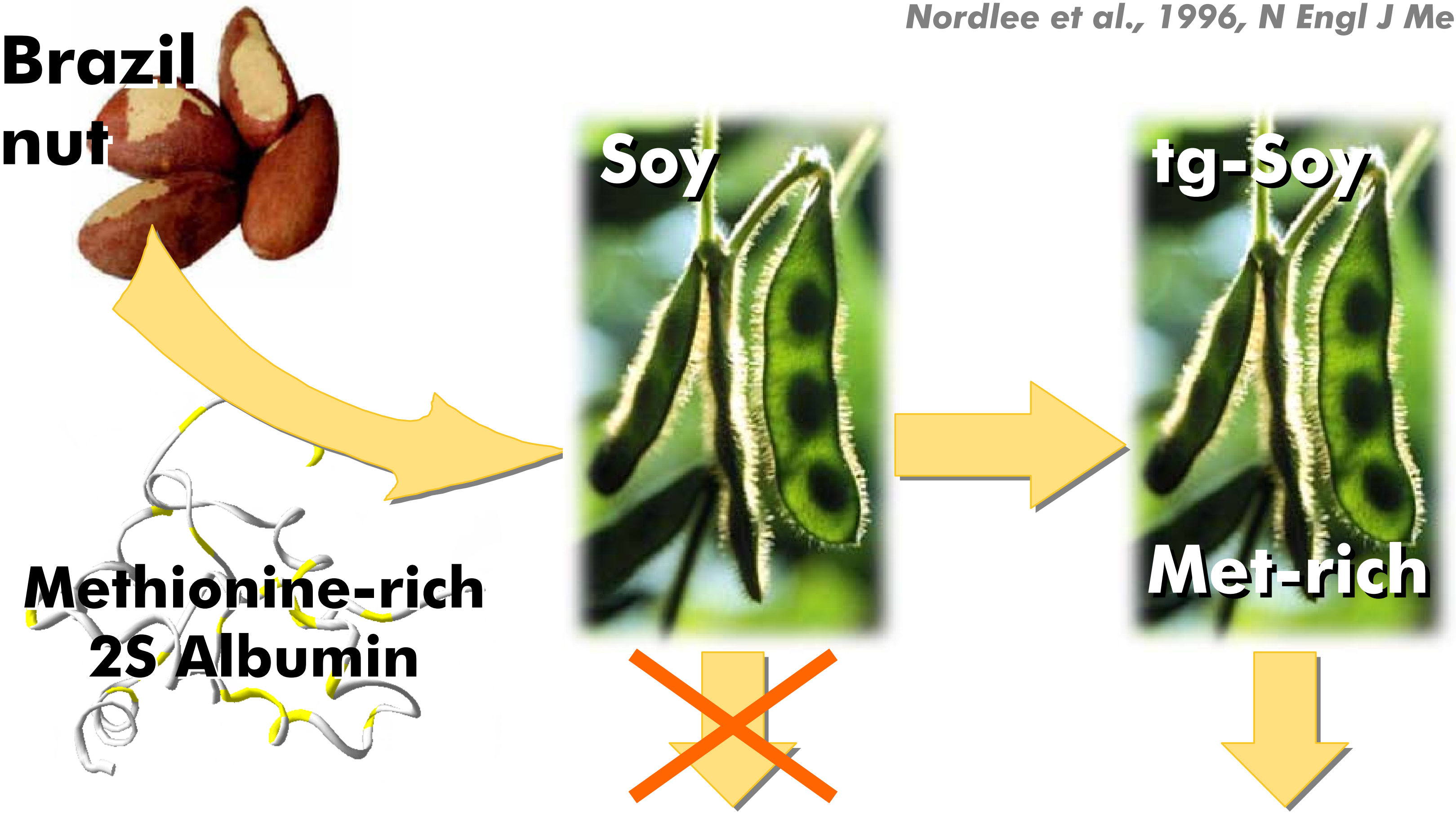
Database Entry	Profile 6		Profile 7		Profile 8		Profile 9		Profile 10	
	weight	score	weight	score	weight	score	weight	score	weight	score
CDS1_SCHPO		5.5		7.0		7.3		7.9		7.8
FHL1_YEAST	8.5	20.4	6.5	17.5	6.6	17.9	6.2	15.9	6.0	16.4
FKH1_YEAST	9.6	23.7	7.1	19.2	6.3	18.7	5.9	16.6	6.0	16.7
FKH2_YEAST	8.8	22.7	6.6	18.4	6.2	18.3	5.7	15.9	5.6	16.2
FRAH_ANASP		9.6	6.3	16.0	6.0	17.1	5.7	15.4	5.6	15.8
KAPP_ARATH		5.3		7.2		8.2		10.6	6.0	12.9
KI67_HUMAN		3.7		5.0		5.4		5.8		5.7
MNF_MOUSE	10.1	18.1	7.4	15.2	7.6	15.6	7.0	13.8	6.8	13.8
SPK1_YEAST		6.2		6.7		7.0		8.1		7.8
YHR5_YEAST	9.2	21.1	6.8	19.2	6.1	19.6	5.6	17.2	5.7	17.8
YKI5_CAEEL		8.5		10.3	8.4	15.0	7.8	13.6	7.6	14.0
HUMKIAA10_1		5.4		8.1		8.4		8.4		8.4
MLB1770_16		6.3		8.5		10.0	7.5	13.5	6.4	13.9
SC9346_10		10.2	7.0	17.2	5.7	17.5	5.4	15.4	5.0	15.9
SCCXIV38K_16	9.7	20.8	6.7	19.1	6.0	19.4	5.5	17.1	5.4	17.9
SCD9717_7	9.7	16.4	7.5	13.9	7.7	14.2	7.1	12.6	6.6	12.7
SCPPR1GEN_4		4.7		7.1		7.8		7.6		8.0
SPAC17G8_10	11.1	17.8	8.7	17.4	6.5	18.5	6.0	16.3	5.5	17.4
SYCSLRG_6		9.1	6.4	15.6	6.1	16.7	5.2	15.5	5.3	15.7
SYCSLRG_63	13.2	15.7	8.7	13.5	7.7	13.7	7.0	13.8	4.9	13.8
YSCL8083_15		9.6	7.0	16.8	6.0	17.1	5.9	14.8	5.4	15.5
YSCL9470_15	10.1	18.7	7.3	15.7	7.2	16.3	6.5	14.3	6.2	14.2
B61188		3.7		5.3		6.1		7.2		7.5
Highest false positive		7.0		7.2		7.4		6.9		7.0

Overview

- Allergy
  - What is an allergen/allergic reaction?
- Generalized profiles
  - Modeling sequence motifs
  - Construction of a profile
- Allergenicity prediction
  - Current prediction algorithm (FAO/WHO)
  - Profile-based prediction

Allergens in Transgenic Food

Nordlee et al., 1996, N Engl J Med



Brazil nut

Methionine-rich 2S Albumin

Soy

tg-Soy

Met-rich

allergic reaction in nut-sensitized patients

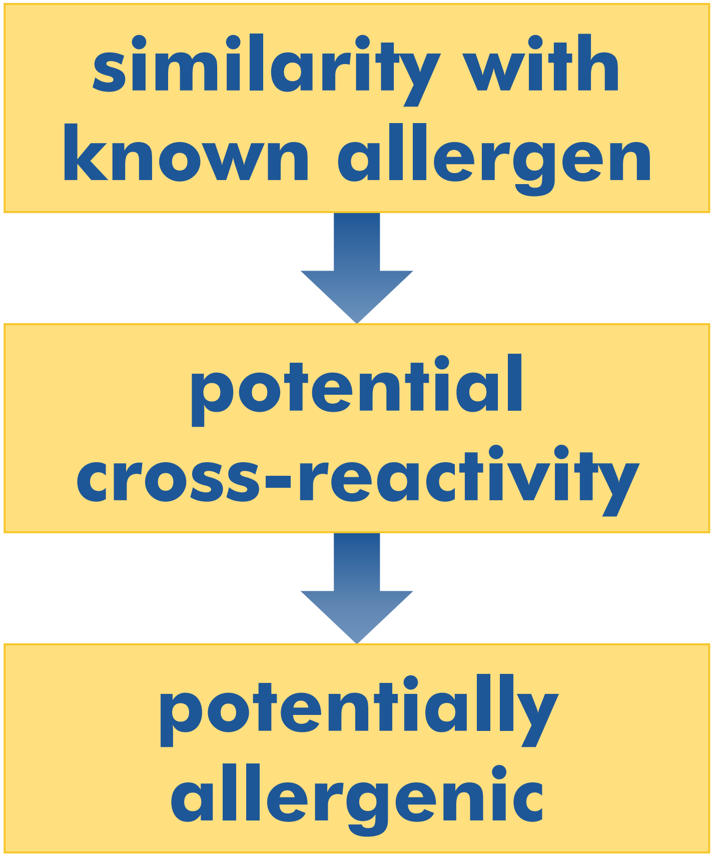
Prediction of Allergens

No common:

- Structure
- Biochemical properties

⇒ no direct prediction

indirect prediction:

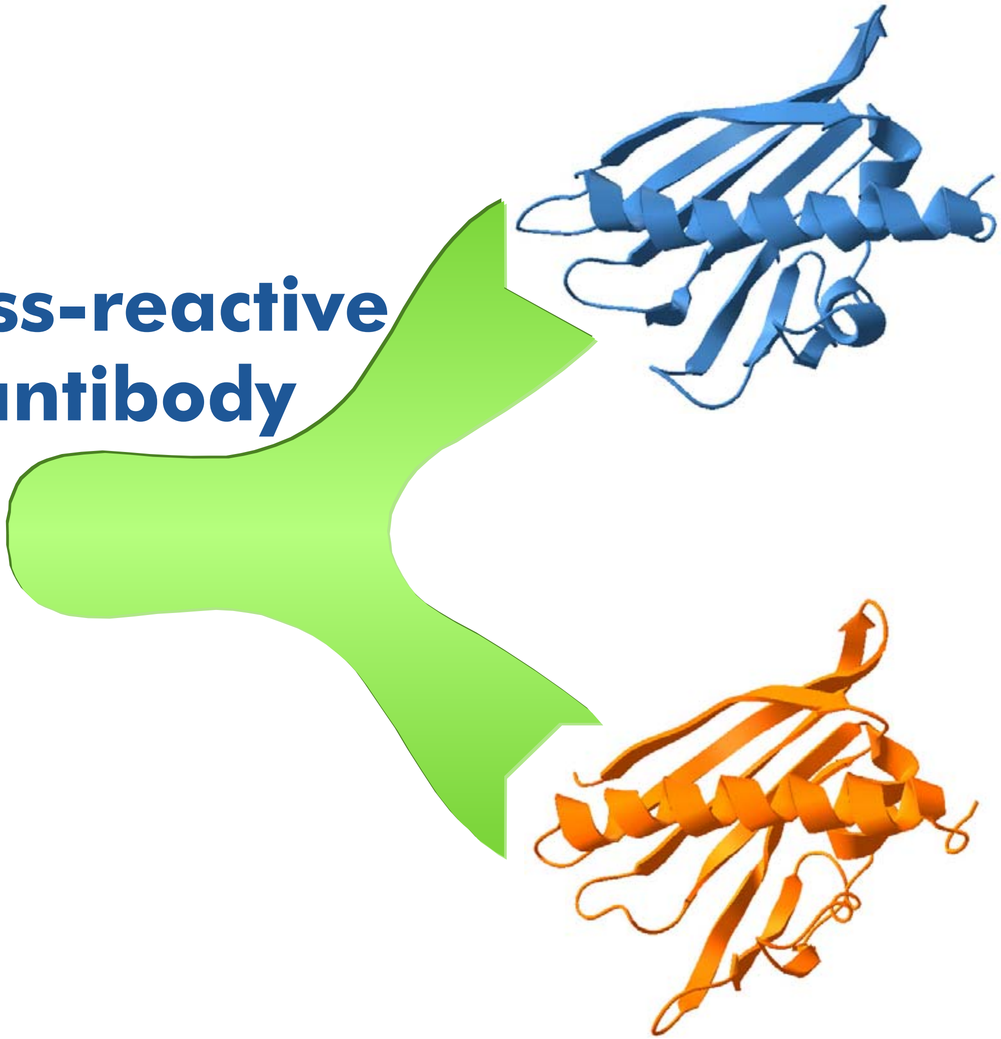


similarity with known allergen

potential cross-reactivity

potentially allergenic

Immunological Cross-reactivity



cross-reactive antibody

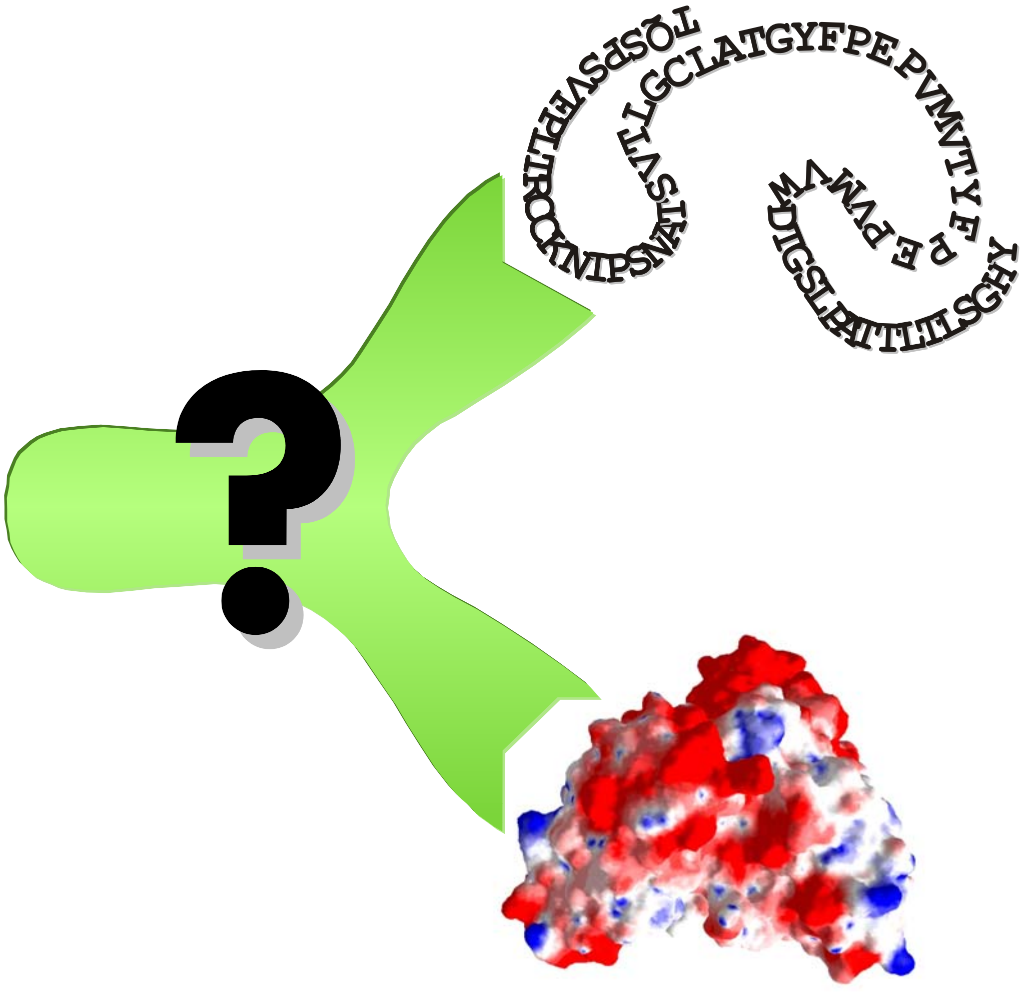
antigen A

birch pollen allergen
(Bet v 1, PDB:1BTV)

antigen B

cherry allergen
(Pru av 1, PDB:1E09)

Prediction of Cross-reactivity



‚linear‘ sequence

- single loops
- secondary structure elements

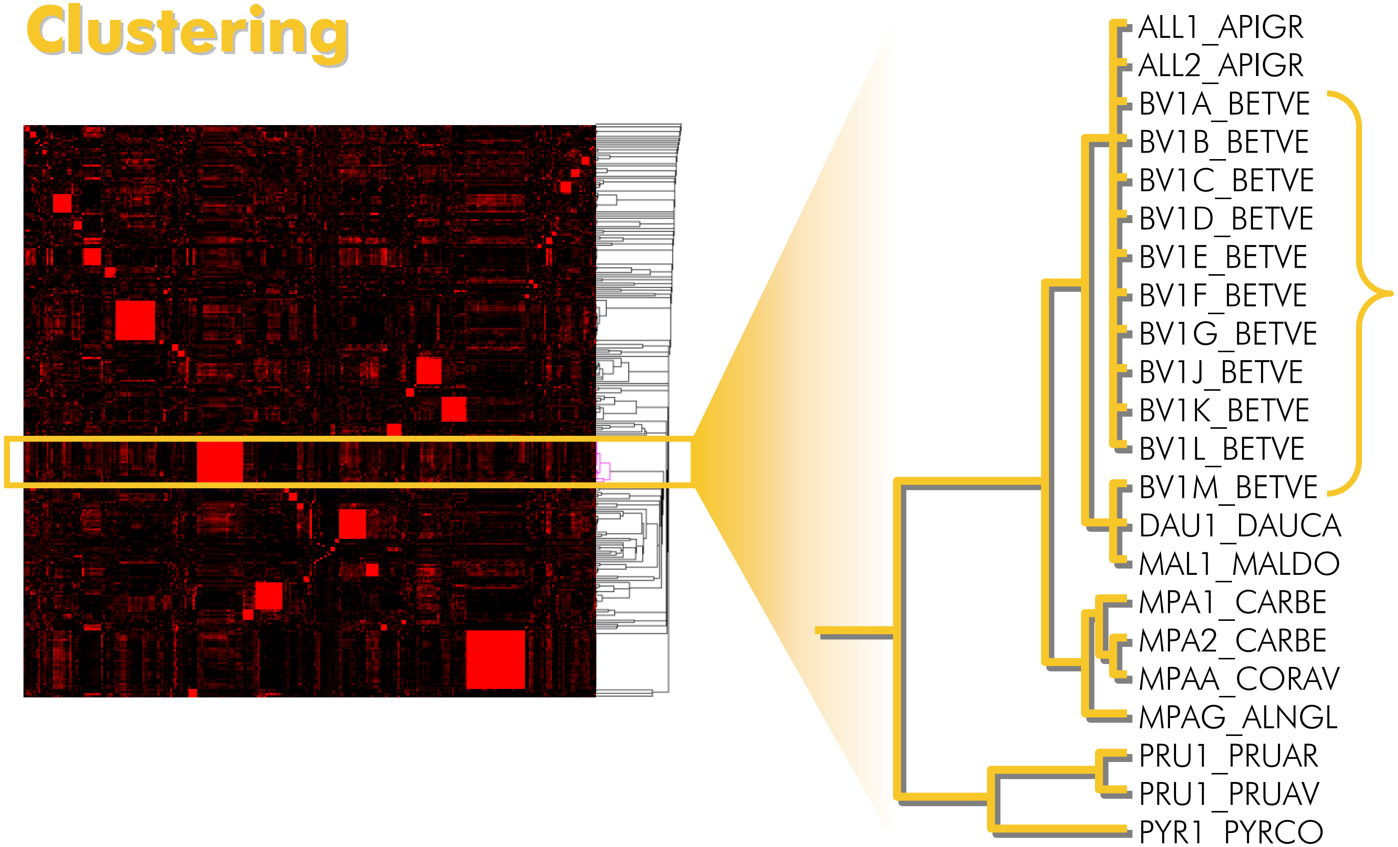
surface

- shape
- physico-chemical properties

Cross-reactivity and Sequence Similarity

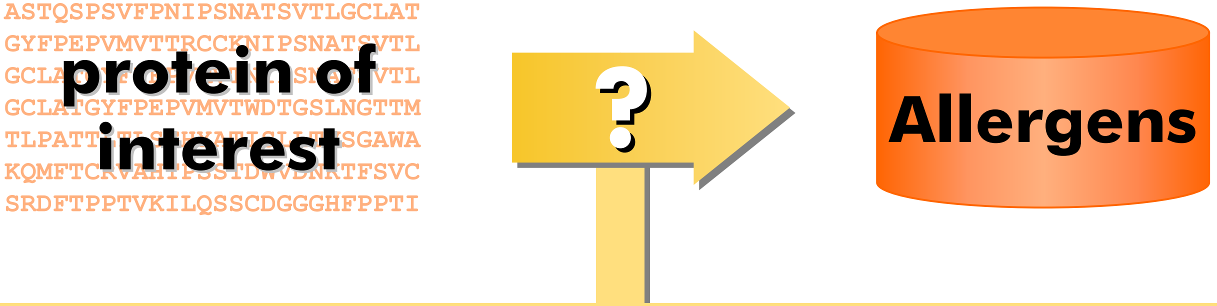
Sequence-based Clustering

Bet v 1 - cluster



ALL1_APIGR	Api g 1	celery
ALL2_APIGR	Api g 2	
BV1A_BETVE	Bet v 1	birch
BV1B_BETVE		
BV1C_BETVE		
BV1D_BETVE		
BV1E_BETVE		
BV1F_BETVE		
BV1G_BETVE		
BV1J_BETVE		
BV1K_BETVE		
BV1L_BETVE		
BV1M_BETVE		
DAU1_DAUCA	Dau c 1	carrot
MAL1_MALDO	Mal d 1	apple
MPA1_CARBE	Car b 1	hornbeam
MPA2_CARBE	Car b 1	
MPAA_CORAV	Cor a 1	hazelnut
MPAG_ALNGL	Aln g 1	alder
PRU1_PRUAR	Pru ar 1	apricot
PRU1_PRUAV	Pru av 1	cherry
PYR1_PYRCO	Pyr c 1	pear

Proposed Allergenicity Evaluation (WHO/FAO)


ASTQSPSVFPNIPSNATSVTLGCLAT
GYFPEPVMVTTRCCKNIPSNATSVTL
GCLATGYFPEPVMVTWDTGSLNGTTM
KQMFTCRVAHTPSSTDWVDNKTFSVC
SRDFTPPTVKILQSSCDGGGHFPPTI
protein of interest
?
Allergens

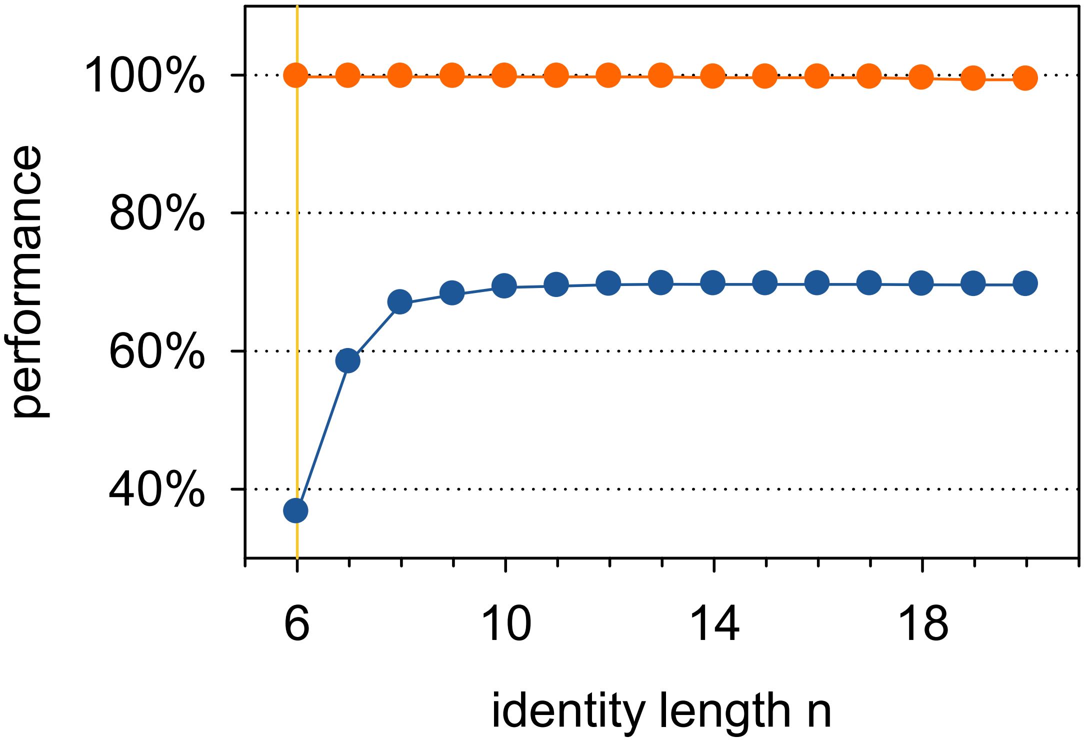
- **identity-test:**
  n = 6 contiguous amino acids
- **similarity-test:**
  35 % (80 residues)

Allergen Prediction According to Guidelines

Database:	**Adb**	**Swiss-Prot**	**Rice**	**trGEN (human)**
Description:	allergen database	general protein database	rice genome (TIGR OsGI)	human genome (translated)
#Proteins:	**779**	**101’602**	**10’891**	**330’743**
allergens (predicted)	**98.6 %**	**67.3 %**	**75.9 %**	**42.9 %**

clinical observation: < 0.5 %

Performance of Current Allergen Prediction

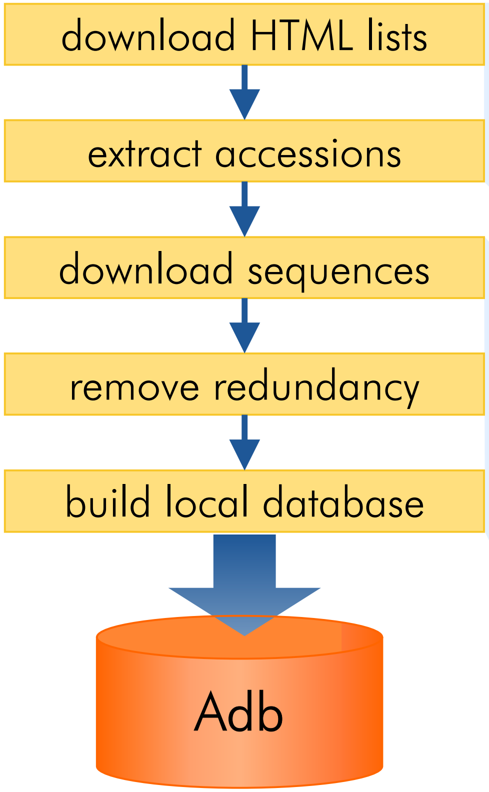


recall:
% true allergens predicted to be an allergen

precision:
% true allergens of predicted allergens

unsuitable for allergenicity prediction.

Allergen Database (Adb)



download HTML lists

extract accessions

download sequences

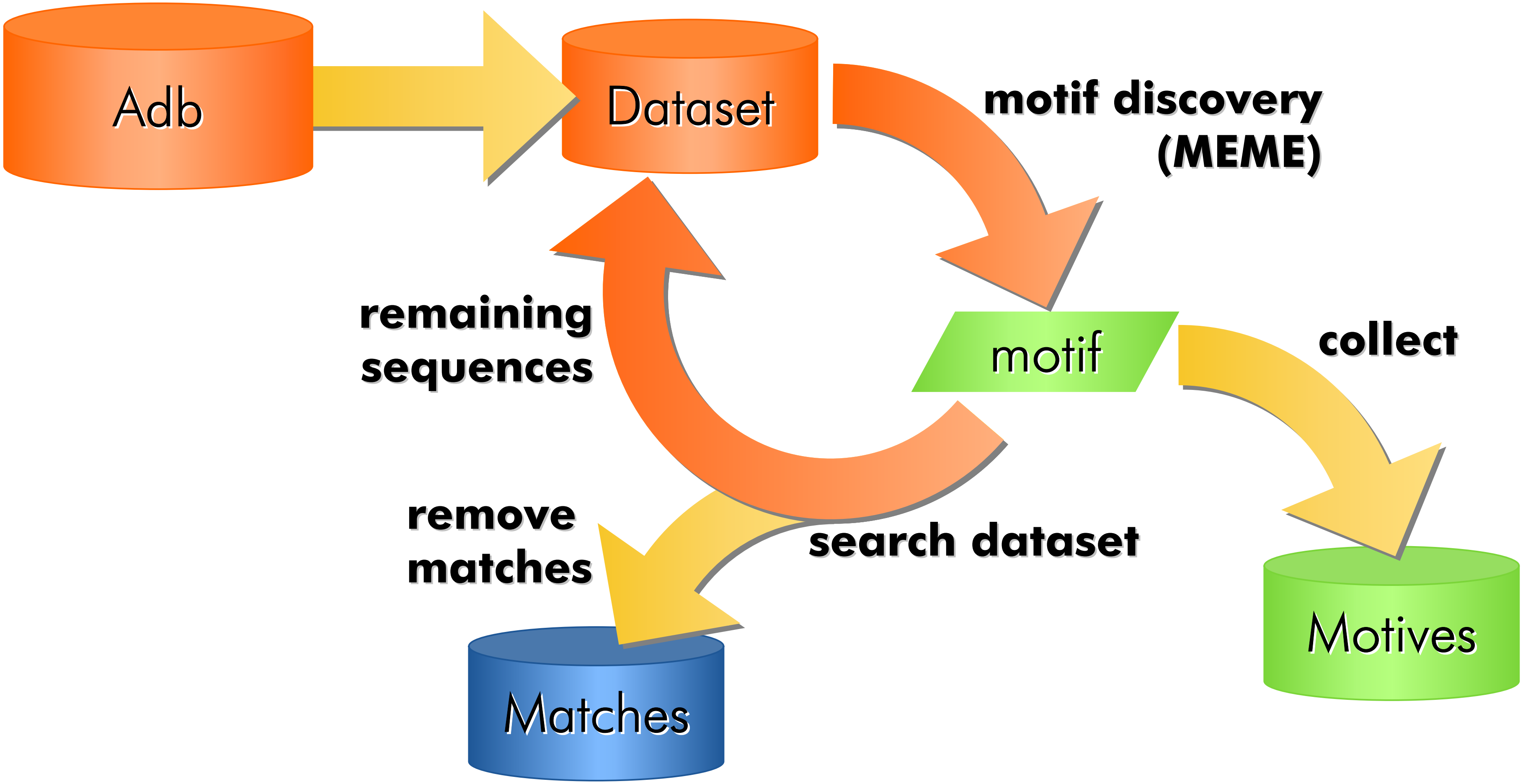
remove redundancy

build local database

Data source (February 11th, 2002)	#accessions
www.allergen.org (isoallergens)	165
www.expasy.org (Swiss-Prot allergens)	276
www.iit.edu/~sgendel (food allergens)	140
www.iit.edu/~sgendel (non-food allergens)	341
add_list.txt	105
Total accessions:	809

Database		#sequences
Swiss-Prot		439
Swiss-Prot (variants)		99
GenBank		269
PIR		97
skipped:	redundant	-111
	unreachable	-7
	annotation problems	-7
Total retrieved sequences:		779

Automatic Identification of Allergen Profiles



Adb

Dataset

motif discovery (MEME)

motif

collect

Motives

remaining sequences

search dataset

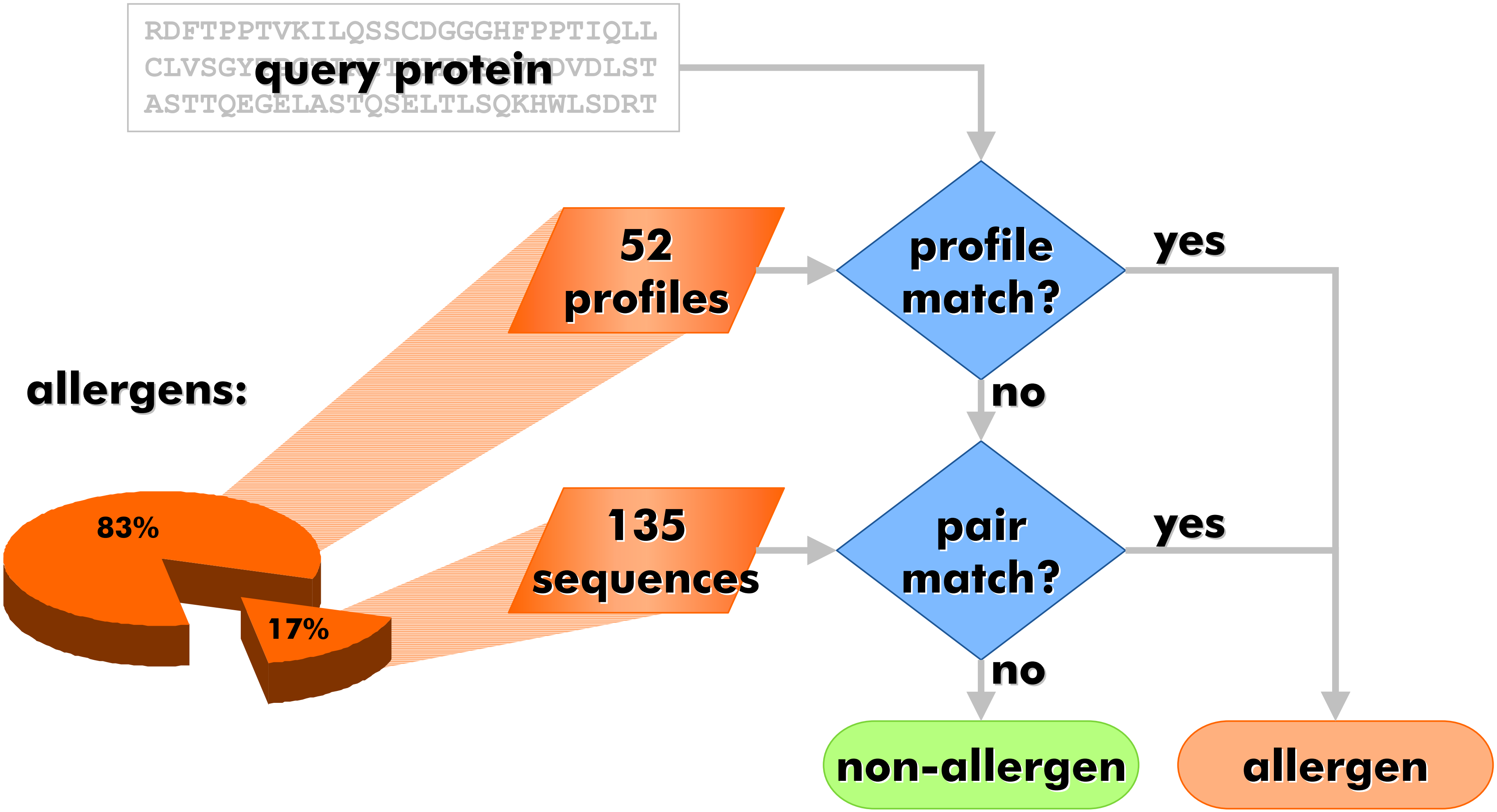
remove matches

Matches

52 Allergen Profiles

Profile Identifier	*MEME E-value*	*Matching allergens*	*Predominant PROSITE matches*
AM00001	**$1.8 \cdot 10^{-4123}$**	**101**	**Pathogenesis-related Bet v 1 family**
AM00002	**$2.0 \cdot 10^{-1477}$**	**68**	**Profilins Pollen proteins (Ole e I)**
AM00003	**$1.3 \cdot 10^{-919}$**	**36**	**Globins**
AM00004	**$3.0 \cdot 10^{-845}$**	**35**	**none**
AM00005	**$4.8 \cdot 10^{-794}$**	**22**	**SCP/Tpx-1/Ag5/PR-1/Sc7**
...	**...**	**...**	**...**

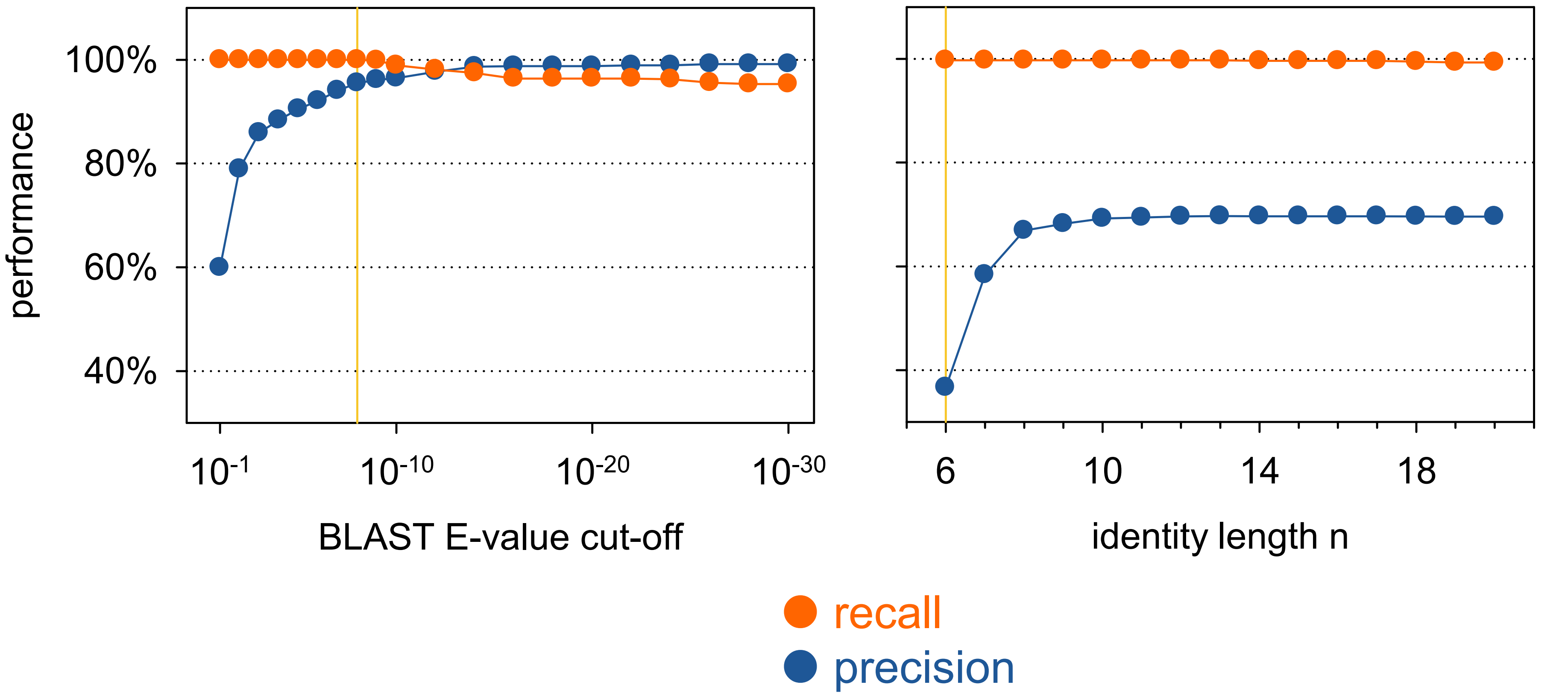
Profile-based Allergenicity Prediction



Performance of Prediction: Random Sequences

Profile-based

FAO/WHO



Performance of Prediction: Swiss-Prot Sequences

Prediction method	*Predicted Allergens*	*True Allergens*
FAO/WHO	68’356 67.3 %	351 0.5 %
Profile-based	4’096 4.0 %	351 8.6 %

Conclusion

- **Allergen Database (Adb):**
  - most complete allergen sequence resource
  - semi-automated
- **WHO/FAO method for allergenicity evaluation:**
  - unspecific (precision = 0.5%)
  - algorithmic limitation
- **Profile-based prediction:**
  - improved performance (precision = 8.6%)

Outlook

- **Improved prediction:**
  - pure profile based approach
  - two phase prediction (fold, surface)
- **Search for unknown allergens:**
  - xeno- vs. autologous structures
  - production of pan-allergic structures (for diagnosis and therapy)

institute of
immunology
bern

Prof. Dr. B.M. Stadler
Dr. S. Miescher
Dr. M. Vogel
Michaela Fux
Lorenz Scheppler
Tomasz Bobrzynski
Elsbeth Gautschi
Andreas Hofmann
Pamela Marti
Jacqueline de Sa
Diana Arnold
Evelyne Aufdenblatten

Dr. Philipp Bucher
SIB, Lausanne