

C
E
N
T
R
E

F
O
I
R
O
I
N
F
O

G
R
M
A
T
T
I
V
E

V
U

Walter Pirovano
21 sep 2007

- Observed occurrence of amino acids at some position in an alignment that deviates from expected may indicate some (functional) significance

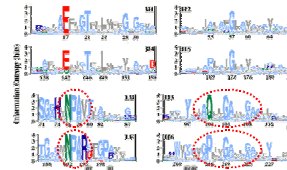
- What 'deviates from expected'?
 - unlikely occurrences
- What is unlikely?
 - only (relatively) few possibilities to obtain observed result

[2] 21 sep 2007 *Water Pirovano*

Pfam Ig Family Alignment

- NPA: stabilizes loops B and E
- G (a) xxxG (a) xxG (a) :

- Crossing of right-hand helical bundles



Andreas Engel and Henning Stahlberg, in: *Current Topics in Membranes* (2001), Hohmann, Agre & Nielsen (Eds.) Academic Press

- Number of possibilities for finding some combination of aminoacids:
 - which types?
 - how much of each?

- WWW 3 W $\Rightarrow$ only 1 way
- RHH 1 R, 2 H $\Rightarrow$ three ways
- SHQ 1 S, 1 H, 1 Q $\Rightarrow$ six ways

[5] 21 sep 2007 Walter Pirovano

- 'Real' examples:

- WWWWWWWWWWWWWWWWWWWWWWWWWWWW
- 33 W ⇒ only 1 way
- RRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRR
- 16 R, 17 H ⇒ ? ways ($\sim 2^{33} \approx 10^9$)
- SSSSHSSCCCCCCCCCEEQQEEEEEEEEEQEE
- 7 S, 1 H, 8 C, 14 E, 3 Q ⇒ ??? ways ($\sim 5^{32} \approx 10^{23}$)

- 'many' ways
- ⇒ but, we can calculate that!

[6] 21 sep 2007 Walter Pirovano

Shannon's 'Information Entropy':

- 'A Mathematical Theory of Communication',
The Bell System Technical Journal, Vol. 27, 1948.

" Can we define a quantity which will measure, in some sense, how much information is 'produced' by such a process, or better, at what rate information is produced? "

- He was thinking about the *Transmission of Information*, i.e., from a Source through some Channel to a Destination.

7| 21 sep 2007 Walter Pirovano



Solution: Entropy

$$H = \sum_{i=1}^n p_i \log p_i$$

- the entropy of a set of probabilities p_i
- measures information, choice and uncertainty
- zero only if only one p_i is not zero
 - there is only one choice
- maximal if all p_i are equal
 - most 'uncertain' situation: all options are possible

8| 21 sep 2007 Walter Pirovano



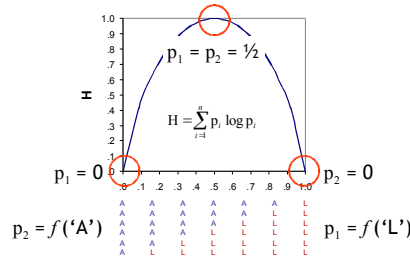
Information Content

- Shannon was thinking about the *Transmission of Information*, i.e., from a Source through some Channel to a Destination.
- ...but it applies equally well to any type of 'message'
- We can use it to measure the level of conservation in columns in an alignment

9| 21 sep 2007 Walter Pirovano



Simple Example: Sequence Entropy



10| 21 sep 2007 Walter Pirovano



Sequence Analysis: Comparing Groups

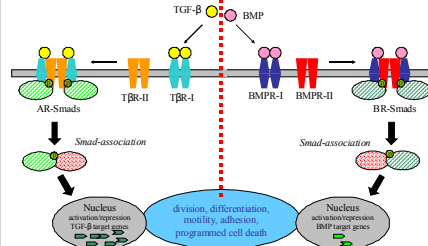
- Many biological problems relate to questions like:
" Why do these proteins do this, and those proteins not? "
- or
" Why do these patients get sick, and those not? "

- ⇒ The answer can be related to similarities and differences between sequences
- Similarities (conservation) relate to functionally critical positions
 - Differences can explain functional differences

11| 21 sep 2007 Walter Pirovano

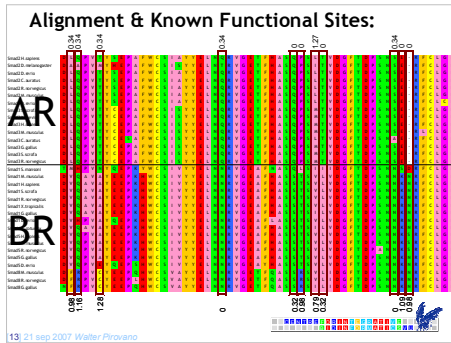


TGF-β signalling pathway



12| 21 sep 2007 Walter Pirovano





Measuring Overlapping Distributions

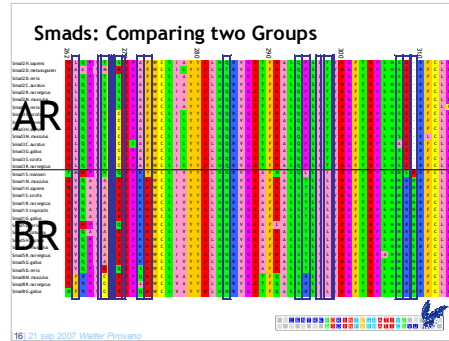
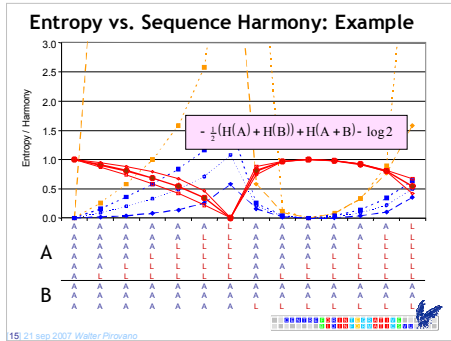
- Weigh both groups equally;
take $p^A + p^B$ in stead of p^{AB} :

$$SH_i^{A/B} = \sum_x p_{i,x}^A \log \frac{p_{i,x}^A}{p_{i,x}^A + p_{i,x}^B}$$

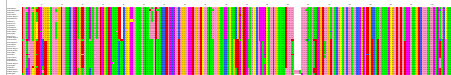
- Fixed interval [0,1], but not completely symmetrical



14: 21 sep 2007 Walter Pirovano



Smad-MH2 Alignment & Functionally Specific Sites



- 29 known sites of functional specificity
- based mostly on site-specific mutants and characterized on affinity for binding to BMPR-I vs. TBR-I receptor types

Method	Predict	Specificity	Error
AMAS	6	21%	3%
TreeDet	21	52%	21%
SDPpred	12	31%	10%



17: 21 sep 2007 Walter Pirovano

Finding Low-harmony sites in Smad-MH2

Pos.	Sec. str.	SH	AR	BR	Interaction
L263	B1'	0	La	Vfa	SARA
T267	B1'	0	Ta	Acad	SARA
S269	loop	0	CS	ES	SH-Smad3
A272	loop	0	A	ES	ES-Smad3
F273	loop	0	F	Hy	SH-Smad3
Q284	B2	0	QE	N	TGFβ1
Q284	loop	0.16	Q	Sq	c-Smad3/Smad1
P295	B3	0	P	Tr1	c-Smad3/Smad1
L297	B3	0.11	LHG	Vt	c-Smad3/Smad1
T298	B3	0	T	LS	c-Smad3/Smad1
S308	L1	0	Sa	N	c-Smad3/Smad1
-	L1	0	-	Ned	c-Smad3/Smad1
E309	L1	0	E	Krs	c-Smad3/Smad1
A323	HR	0	Aa	S	Alx12
Y325	HR	0	Y	T	SH-Smad3
M327	HR	0	LRQ	W	Alx12
R334	Loop	0.15	RS	K	SH-Smad3
R337	BS	0	R	H	SH-Smad3

Method	Predict	Specificity	Error
AMAS	6	21%	3%
TreeDet	21	52%	21%
SDPpred	12	31%	10%
Sequence Harmony	(same) 32 (diff=4) 30	75% 93%	28% 33%

Pirovano, Fumetris & Heringer. "Sequence Comparisons by Sequence Harmony Identifies Subgroup-Specific Functional Sites", Nucleic Acids Res., in press (2006).
www.fim.uni-erlangen.de/rohr/2006/Sequence_Harmony.pdf



18: 21 sep 2007 Walter Pirovano

[19] 21 sep 2007 *Walter Pirovano*

- 40 Sites of Low Sequence Harmony in Smad-MH2
 - different between the AR (TGF- β) and BR (BMP) sub-type Smads
 - Low Harmony sites in Smad-MH2 are functionally relevant
 - Other methods cannot select all known sites!
- ⇒ Functional Sites are Interaction Surfaces on Protein Surface:
 - ⇒ Next: Analyze Interaction Partners in the Pathway
- 14 Low Harmony Sites in Smad-MH2 of unknown function
 - 11 putative functions from structural considerations
 - promising candidates that determine TGF- β /BMP specificity
 - confirm (or rebuke) putative functions?

© 21 sep 2007 Walter Pavlou

[20] 21 sep 2007 Walter Pirovano

[illegible]

[21] 21 sep 2007 *Walter Pirovano*

[illegible]

[22] 21 sep 2007 Walter Pirovano

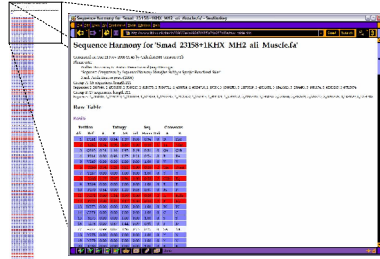
[illegible]

[23] 21 sep 2007 Walter Pirovano

The screenshot shows the 'Import' window of the 'Chemical Shifts' software. The 'Reference Structure (optional):' dialog box is open, with the 'NMR' tab selected. The 'NMR' tab displays a table of chemical shifts with columns for 'Chemical Shift (ppm)' and 'Assignment'. The 'PDF' tab is also visible, showing a similar table. The 'Submit' button is highlighted in blue.

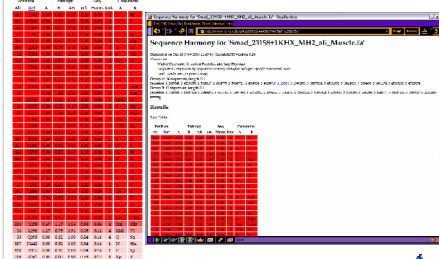
[24] 21 sep 2007 Walter Pirovano

SMAD Sequence Harmony: Raw Table



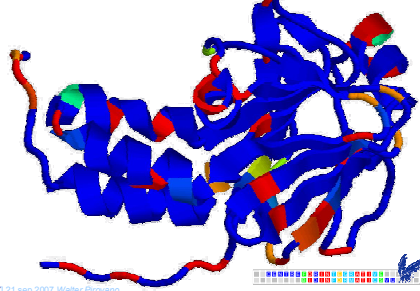
[25] 21 sep 2007 *Walter Pirovano*

SMAD Sequence Harmony: Results



26] 21 sep 2007 Walter Pirovano

SMAD Sequence Harmony: Structure



[27] 21 sep 2007 Walter Pirovano