



# Sequence Entropy

## Genome Analysis

*Walter Pirovano*  
21 sep 2007

# Significance of Alignment Positions

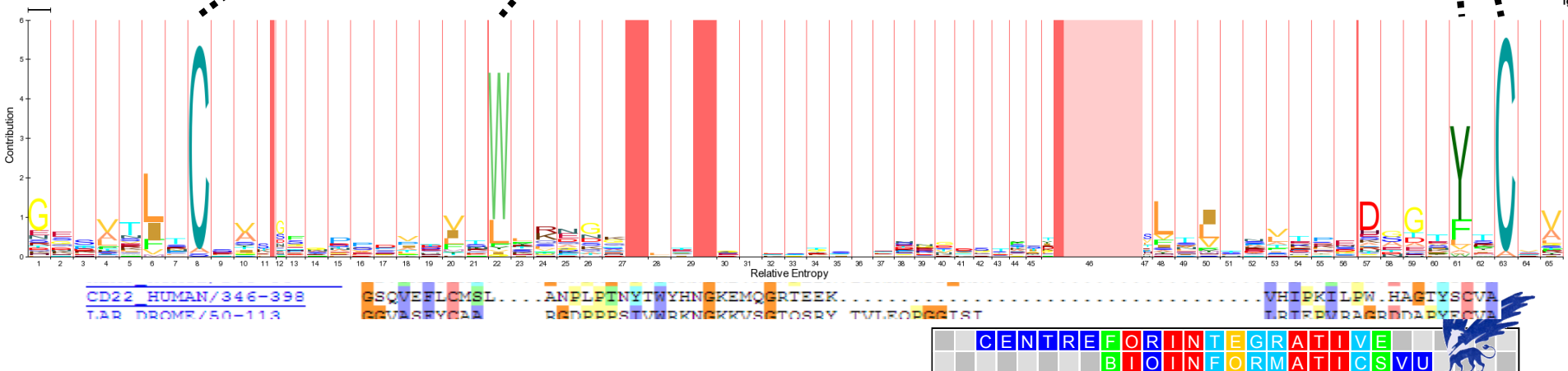
- Observed occurrence of amino acids at some position in an alignment that deviates from expected may indicate some (functional) significance
- What 'deviates from expected'?
  - unlikely occurrences
- What is unlikely?
  - only (relatively) few possibilities to obtain observed result



# Pfam Ig Family Alignment

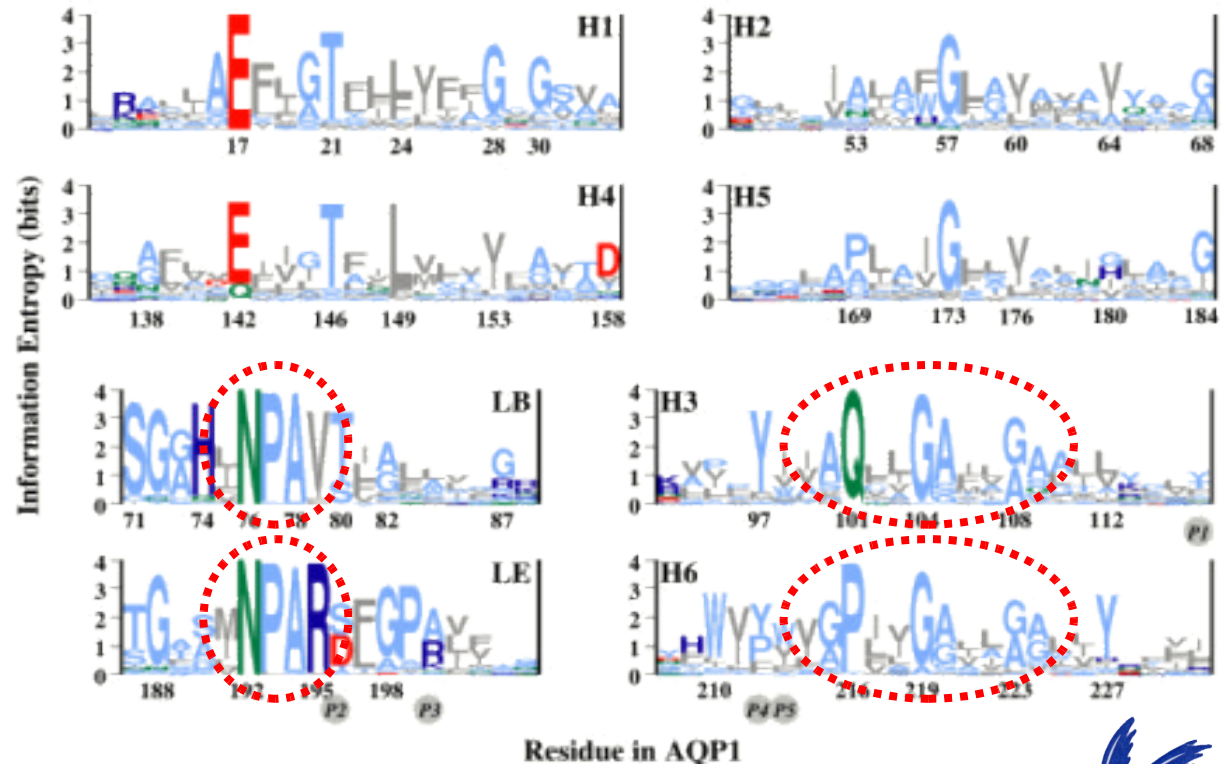
Q31499 ONCMY/211-278  
 1A01 GORGO/220-285  
 Q31463 MORSA/126-191  
 Q31377 CYPCA/123-188  
 Q31340 AGEPH/114-179  
 Q31312 APHCE/113-178  
 HB21 HUMAN/142-207  
 Q31362 BRARE/125-190  
 2DMB HUMAN/128-194  
 HA21 HUMAN/126-191  
 IGHD RAT/93-162  
 2DMA HUMAN/140-204  
 B2MG BRAAR/38-102  
 PVR CERAE/259-314  
 HVC1 HETFR/160-223  
 IGHA1 GORGO/243-315  
 MUCB HUMAN/282-353  
 GCBM MOUSE/248-315  
 IGHE MOUSE/219-287  
 CD48 MOUSE/144-195  
 IGHI1M MOUSE/20-84  
 O57349 CHICK/131-195  
 O46819 TRISC/205-269  
 MUCM RABIT/21-92  
 PTP2 CAEL/62-132  
 P79988 CHICK/35-116  
 SEM3E MOUSE/595-656  
 FCGR1 MOUSE/214-271  
 CEAM1 HUMAN/341-398  
 CD4 MOUSE/35-114  
 OPCM RAT/237-298  
 PECA1 HUMAN/340-398  
 FAS2 SCHAM/145-205  
 MY19 CHICK/131-195

TPSSPVTCCHAI...GFYPSCGVWVFWQKDGQEQHGDVEHC...EILQNDGDTFQKS...THLTVPPEEWK...NNKYQCQV  
 DHEAILRCWAL...SFYPAEITLTWQRDGEDQTDTELV...STRPAGDGTTFQK...WAAVVVPSG...QEQRVYTCV  
 KHPSMLVCSVY...DFFPKHIRVSWLRDGEVTSVDTST...DELADADWIFYQI...HSHLEYTPR...SGEKISCVV  
 SRPAVLVCSAY...DFYPPKKIKVSWLRDGEKEMTSVDTST...MEMADGDWIFYQI...HSELEYTPK...SGEKISCMV  
 PGPDRLLCSVM...DFYPAAIQVRWFQGGQELSEHVAT...DMVPNGDWTYQL...LVLLLETPR...RGLTYSQV  
 ANPGRLLCSVM...DFYPAEIQLRWFQGGHQLSVHVVT...DIVPNGDWTYQL...LVLLLETPR...LGVSYTCV  
 NHHNLLVCSVT...DFYPAQIKVRWFQGGHQLSVHVVT...DIVPNGDWTYQL...LVLLLETPR...LGVSYTCV  
 DIENKLICHVT...GFFPPFVRVSWIKNNEIVTEGMSVS...QYRPNNDDGTINI...FSTLRFTEV...EGDIYSCSV  
 REPVMLACYVW...GFYPAEVTITWRKNGKLVMPHSSAHKTAQPNGDWTYQT...LSHLALTPS...YGDYTCV  
 GQPNLTICLVD...NIFPPVNNITWLSNGHVSVEGVSET...SFLSKSDHSFFK...ISYLTFLPS...ADEIYDCKV  
 QMSGWLMEVVS...GFYPEDIEHLWLSAQTKMDPINFVIAQFVRQSGDKFQIW...SVLRLEVALSP...SLDTYTCV  
 GKPNTLVCFVS...NLFPMLTVNWHDSVPEVGFCTF...VSAVDGLSFOA...FSYLNFTPE...PSDIEFCIV  
 GKPNTLVCFVS...NLFPMLTVNWHDSVPEVGFCTF...VSAVDGLSFOA...FSYLNFTPE...PSDIEFCIV  
 QNEATLTCDAR...SNPEPTGYNWSTTMGLPFFFAVAQ...GAQ...LLIRFVDEK...INTTFICNV  
 EATVTLTCVVS...NAPYGVNVSVTQEQKPLKSEIAVQ...FGEDPDSVIST...VDISTQAWL...SEAVFYCVV  
 NELVTLTCLAR...GFSPKDVLRWLQGSQELPREKYLT...WASRQEPSQGITFA...VTSILRVAAEDWK...KGDTEFCMV  
 RESATITCLVT...GFSPADVVFQWMQRCQPLSPKEYVT...SAPMPEPQAPERYFA...HSILTUSEEWN...TGETYTCV  
 RKDVSLTCLVV...GFNPGDISVEWTSNGHTEENYKDTA...PVLDSGDSYFIYS...KLNMTKSKWE...KTDSFSCNV  
 NGAPKLTCLVVD...LESEKNVNVWQEKKTSSVSASQWY...TKHHNNATTSITS...ILPVVAKDWI...EGYCYQCIV  
 SCHLRISCEV...KDQHVDTYTYESSGFPPKKSPCY...V...LDLIVTPQN...KSTFYTCV  
 NSMVTILGCLVK...GFYFEPVTVIWNSSLSGCVHTFPA...VLQSDLYTLS...SSVTVPSPPR...PSETVTCNV  
 EGRVTLSCYTQ...GCKPQPQITWLLDNGIQLPGDTRH...KLEADGKKWTTT...STLTVLAYG...PNSTATCLV  
 GDSKKLSCVAT...GFYPPQSIELKLRDQAAADVPSYG...VRPNHGDYQM...EKQTDSEPS...DPAKFSCEV  
 GNLVAMGCLAR...DFLPSSVTFSSWFKNNSEISSRTVR...TFPVVKRGDKYMATS...QVLVPSKDVLCQ...TEEYLVCKV  
 TQLALVNCSE...NGVTFQNYWKLKSHVVDILAD...SSRQWVSDERYQLIR...BFYTLRVSEVTP...DSGTYRCRL  
 NNSTEILCPM...KGEHTEVYWRWNGRQHFEFL...FSSPLGKATYGTNISQEKFSI...RETSSYHSYRLHINRLHS...DNGTYKCT  
 SNSTIECTP...RSLQAKTFVQKGRDVRKEEVK...TDDRVMKMDLG...LLFLRVKRS...DAGTYFCT  
 GSIVTLNCEINLLQRPGLQLSFSFYVSKILEYRNTSS...E...YHIARAERE...DAGTYFGEV  
 FDSVNLTCST...NDTQTSIRWFFKNQSLPSSERMK...LSQGNIT...LSINPVKRE...DAGTYFGEV  
 GESAELEPCS...SQKKITVFIWKFSDQRKILGQHK...GLVIRGGSPSQFDRFDSKKG...AWEKGSF...LIINKLME...DSQTYCAL  
 GQKGLSCER...SAPMAEFQWFKEDTRLATGLDEV...RIENKGRIST...LTFENVSEK...DYGNYCYA  
 GERLNLSCSI...LSAPPANFTIQEDTIVSQTQDFT...KIASKS...DSQTYCIA  
 NETEKIRCRV...SANPPAIVNWMRDGHIVETGDRYV...VEQDQ...LTILNVTEM...DDGTYCRA  
 MY19 CHICK/131-195...SSPDRSSSTSTTSKATYSSKSTV...TSQETLCS



# Aquaporin: Motifs

- **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

# Counting...

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



# Counting... (2)

- ‘Real’ examples:
    - WWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWW  
    • 33 W                                   $\Rightarrow$  only 1 way
    - RRRRRRRRRRRRRRRRRRRHHHHHHHHHHHHHHHHH  
    • 16 R, 17 H                               $\Rightarrow ?$  ways ( $\sim 2^{33} \approx 10^9$ )
    - SSSSSHSSCCCCCCCCEEQEEEEEEEEEEQEEE  
    • 7 S, 1 H, 8 C, 14 E, 3 Q               $\Rightarrow ???$  ways ( $\sim 5^{32} \approx 10^{23}$ )
  - ‘many’ ways
- $\Rightarrow$  but, we can calculate that!



# 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.

# 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

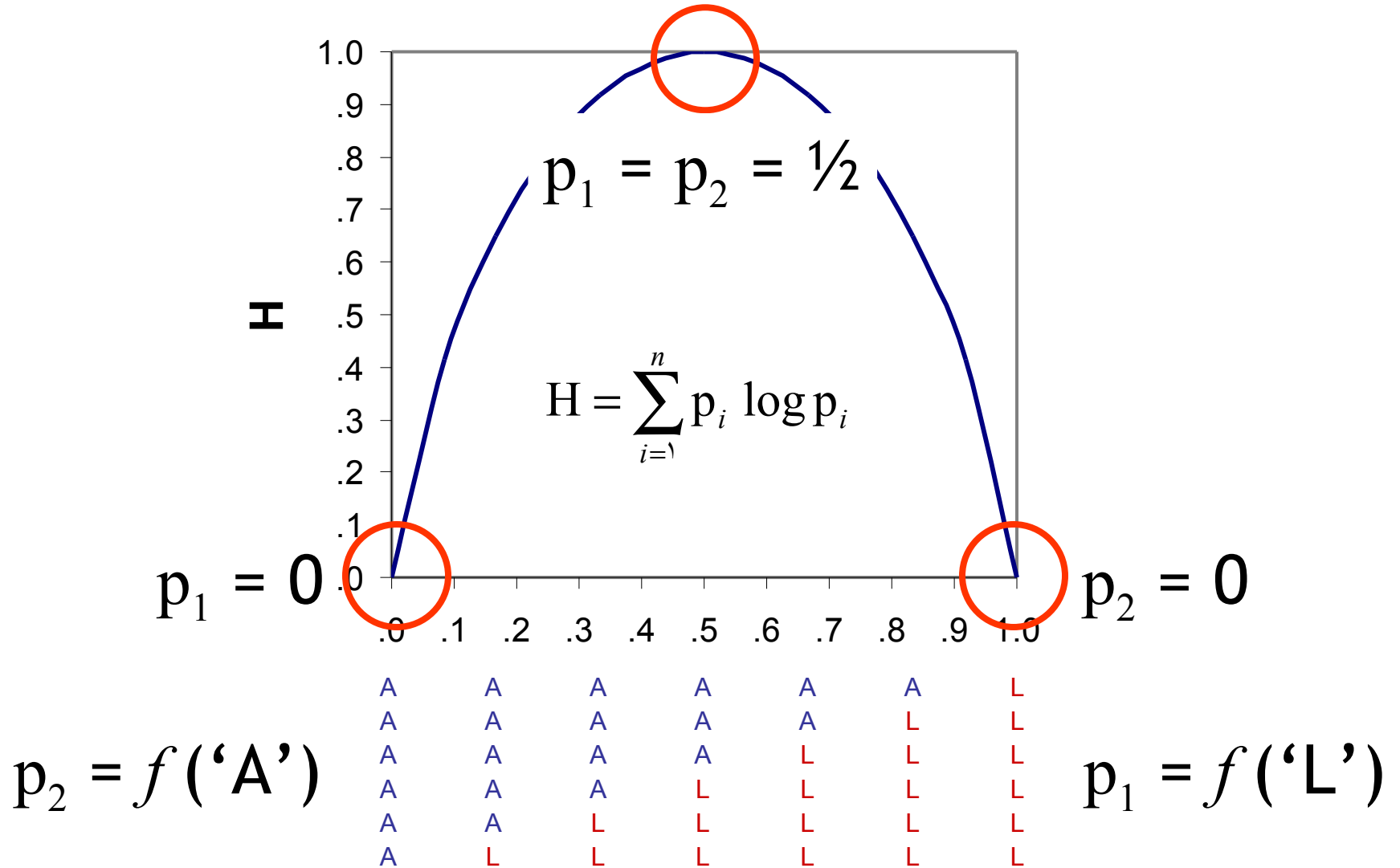


# Information Content

- 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



# Simple Example: Sequence Entropy



# Sequence Analysis: Comparing Groups

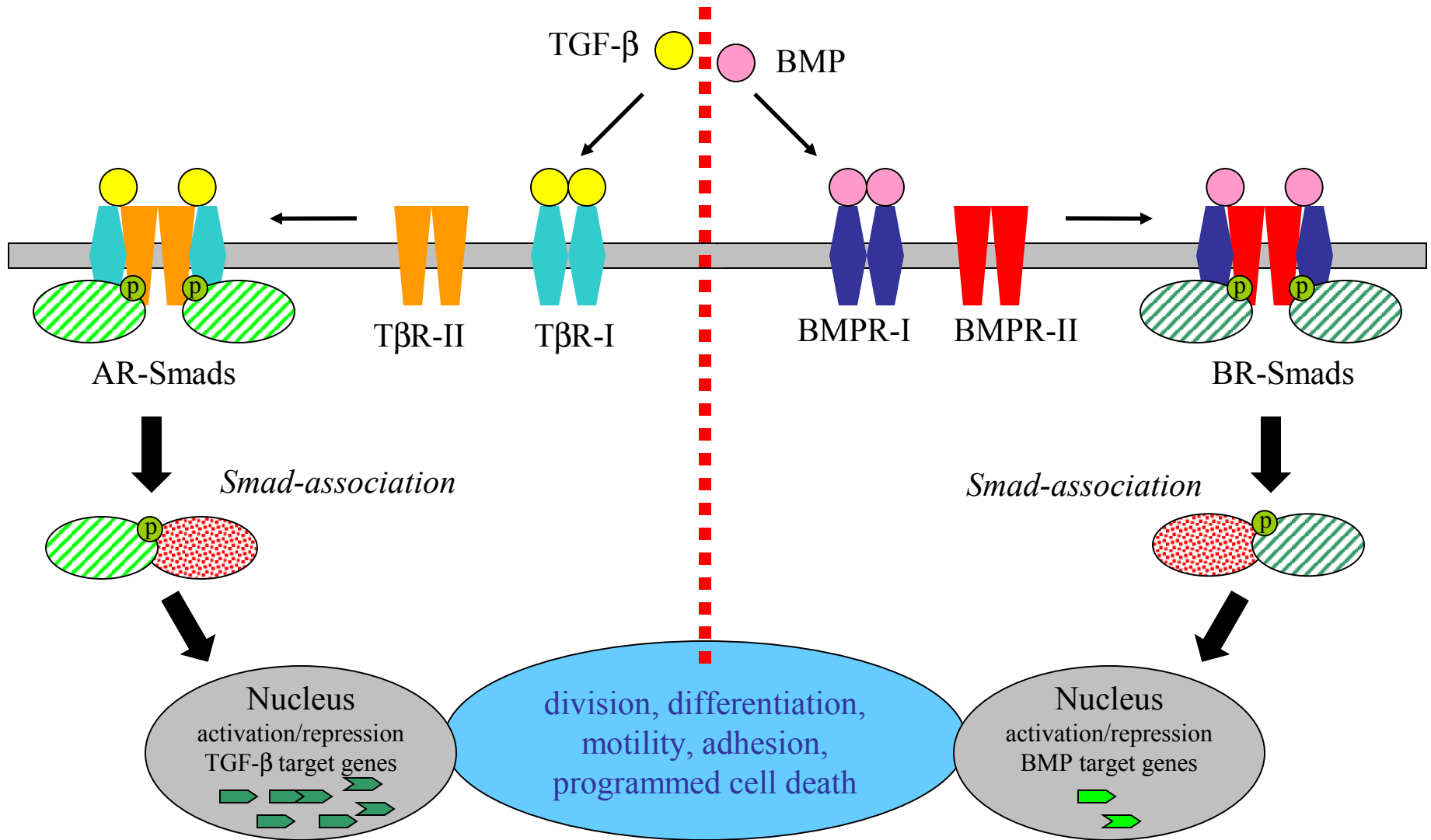
- Many biological problems relate to questions like:  
“ Why do these proteins do this, and those proteins not? ”
- or  
“ Why do these patients get sick, and those not? ”

⇒ The answer can be related to similarities and differences between sequences

- Similarities (conservation) relate to functionally critical positions
- Differences can explain functional differences



# TGF- $\beta$ signalling pathway





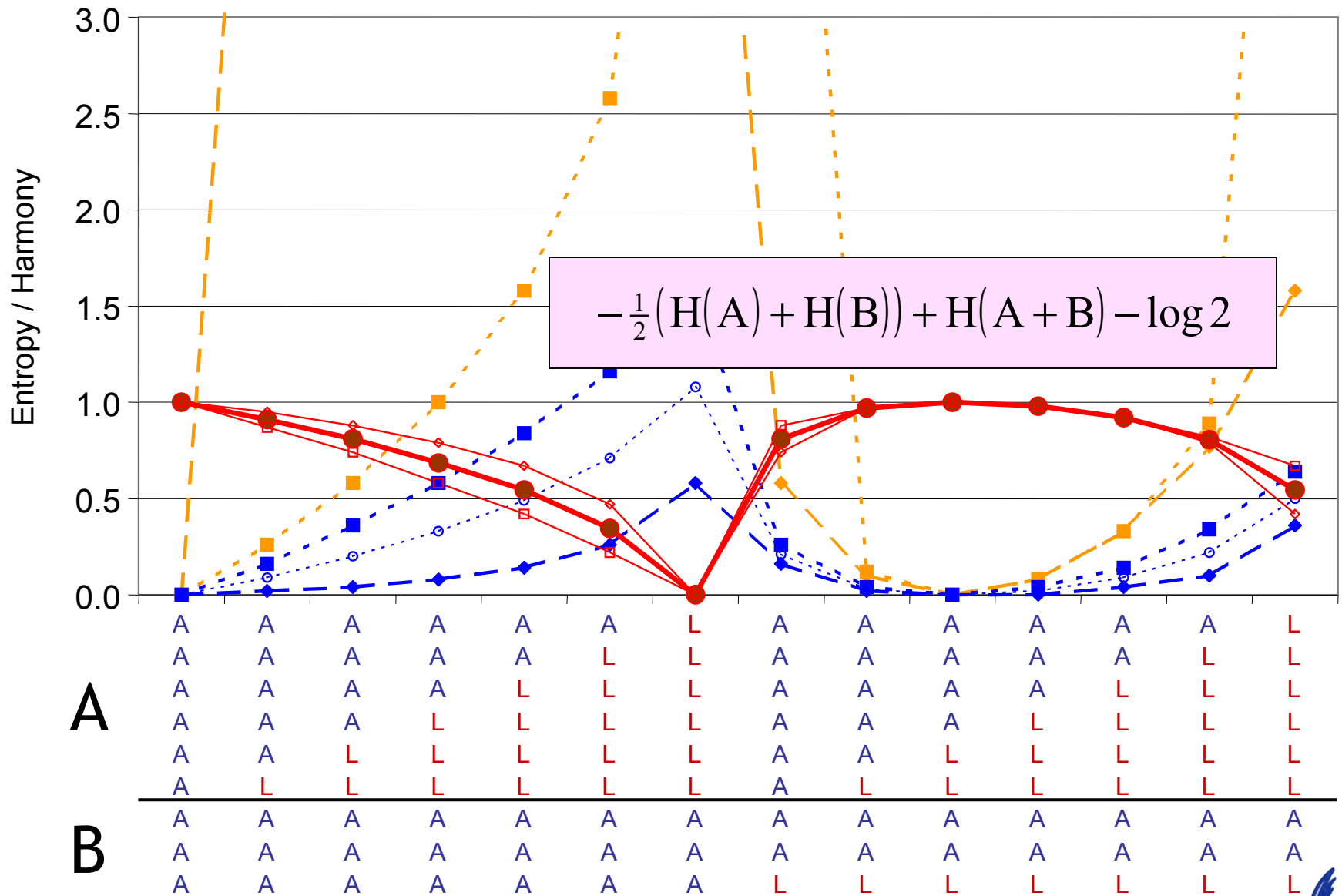
# Measuring Overlapping Distributions

- 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

# Entropy vs. Sequence Harmony: Example



# Smads: Comparing two Groups

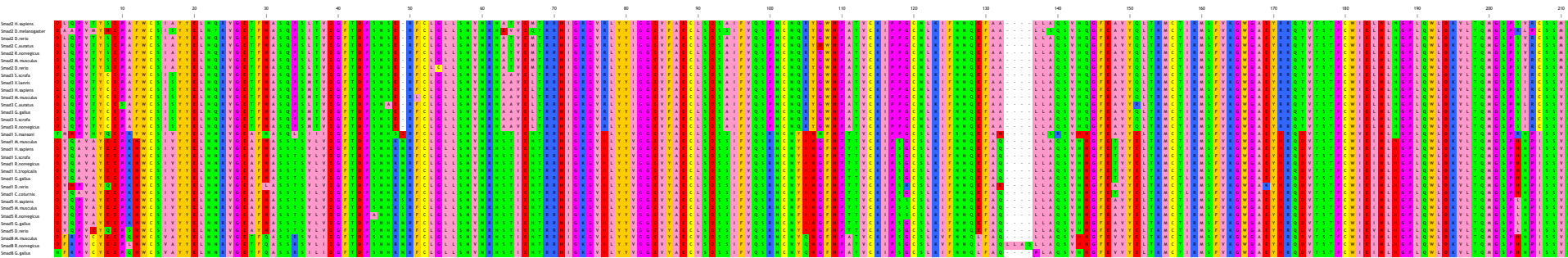
|                      | 262 | 270 | 10 | 20 | 280 | 290 | 30 | 40 | 300 | 310 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|----------------------|-----|-----|----|----|-----|-----|----|----|-----|-----|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Smad2 H.sapiens      | D   | L   | Q  | P  | T   | S   | P  | A  | F   | W   | C | S | I | A | Y | Y | E | L | N | Q | R | V | G | E | T | F | H | A | S | Q | P | S | L | T | D | G | F | T | D | P | S | N | S | E | - | R | F | C | L | G | L |
| Smad2 D.melanogaster | D   | A   | A  | P  | M   | H   | P  | A  | F   | W   | C | S | I | S | Y | Y | E | L | N | T | R | V | G | E | T | F | H | A | S | Q | P | S | L | T | D | G | F | T | D | P | S | N | S | E | - | R | F | C | L | G | L |
| Smad2 D.erio         | D   | L   | Q  | P  | T   | S   | P  | A  | F   | W   | C | S | I | A | Y | Y | E | L | N | Q | R | V | G | E | T | F | H | A | S | Q | P | S | L | T | D | G | F | T | D | P | S | N | S | E | - | R | F | C | L | G | L |
| Smad2 C.auratus      | D   | L   | Q  | P  | T   | S   | P  | A  | F   | W   | C | S | I | A | Y | Y | E | L | N | Q | R | V | G | E | T | F | H | A | S | Q | P | S | L | T | D | G | F | T | D | P | S | N | S | E | - | R | F | C | L | G | L |
| Smad2 R.norvegicus   | D   | L   | Q  | P  | T   | S   | P  | A  | F   | W   | C | S | I | A | Y | Y | E | L | N | Q | R | V | G | E | T | F | H | A | S | Q | P | S | L | T | D | G | F | T | D | P | S | N | S | E | - | R | F | C | L | G | L |
| Smad2 M.musculus     | D   | L   | Q  | P  | T   | S   | P  | A  | F   | W   | C | S | I | A | Y | Y | E | L | N | Q | R | V | G | E | T | F | H | A | S | Q | P | S | L | T | D | G | F | T | D | P | S | N | S | E | - | R | F | C | L | G | L |
| Smad2 D.erio         | D   | L   | Q  | P  | T   | S   | P  | A  | F   | W   | C | S | I | A | Y | Y | E | L | N | Q | R | V | G | E | T | F | H | A | S | Q | P | S | L | T | D | G | F | T | D | P | S | N | S | E | - | R | F | C | L | G | L |
| Smad3 S.scrofa       | D   | L   | Q  | P  | T   | C   | P  | A  | F   | W   | C | S | I | S | Y | Y | E | L | N | Q | R | V | G | E | T | F | H | A | S | Q | P | S | L | T | D | G | F | T | D | P | S | N | S | E | - | R | F | C | L | G | L |
| Smad3 H.sapiens      | D   | L   | Q  | P  | T   | C   | P  | A  | F   | W   | C | S | I | S | Y | Y | E | L | N | Q | R | V | G | E | T | F | H | A | S | Q | P | S | L | T | D | G | F | T | D | P | S | N | S | E | - | R | F | C | L | G | L |
| Smad3 M.musculus     | D   | L   | Q  | P  | T   | C   | P  | A  | F   | W   | C | S | I | S | Y | Y | E | L | N | Q | R | V | G | E | T | F | H | A | S | Q | P | S | L | T | D | G | F | T | D | P | S | N | S | E | - | R | L | C | L | G | L |
| Smad3 C.auratus      | D   | L   | Q  | P  | T   | C   | S  | A  | F   | W   | C | S | I | S | Y | Y | E | L | N | Q | R | V | G | E | T | F | H | A | S | Q | P | S | L | T | D | G | F | T | D | P | S | N | A | E | - | R | F | C | L | G | L |
| Smad3 G.gallus       | D   | L   | Q  | P  | T   | C   | P  | A  | F   | W   | C | S | I | S | Y | Y | E | L | N | Q | R | V | G | E | T | F | H | A | S | Q | P | S | L | T | D | G | F | T | D | P | S | N | S | E | - | R | F | C | L | G | L |
| Smad3 S.scrofa       | D   | L   | Q  | P  | T   | C   | P  | A  | F   | W   | C | S | I | S | Y | Y | E | L | N | Q | R | V | G | E | T | F | H | A | S | Q | P | S | L | T | D | G | F | T | D | P | S | N | S | E | - | R | F | C | L | G | L |
| Smad3 R.norvegicus   | D   | L   | Q  | P  | T   | C   | P  | A  | F   | W   | C | S | I | S | Y | Y | E | L | N | Q | R | V | G | E | T | F | H | A | S | Q | P | S | L | T | D | G | F | T | D | P | S | N | S | E | - | R | F | C | L | G | L |
| Smad1 S.mansoni      | T   | M   | H  | P  | N   | Q   | P  | K  | Y   | W   | C | S | I | V | Y | Y | E | L | N | N | R | V | G | E | A | F | N | A | S | Q | L | S | I | D | G | F | T | D | P | S | N | N | S | D | R | F | C | L | G | L |   |
| Smad1 M.musculus     | D   | V   | Q  | A  | A   | E   | P  | K  | H   | W   | C | S | I | V | Y | Y | E | L | N | N | R | V | G | E | A | F | H | A | S | S | T | S | L | D | G | F | T | D | P | S | N | N | K | N | R | F | C | L | G | L |   |
| Smad1 H.sapiens      | D   | V   | Q  | A  | A   | E   | P  | K  | H   | W   | C | S | I | V | Y | Y | E | L | N | N | R | V | G | E | A | F | H | A | S | S | T | S | L | D | G | F | T | D | P | S | N | N | K | N | R | F | C | L | G | L |   |
| Smad1 S.scrofa       | D   | V   | Q  | A  | A   | E   | P  | K  | H   | W   | C | S | I | V | Y | Y | E | L | N | N | R | V | G | E | A | F | H | A | S | S | T | S | L | D | G | F | T | D | P | S | N | N | K | N | R | F | C | L | G | L |   |
| Smad1 R.norvegicus   | D   | V   | Q  | A  | A   | E   | P  | K  | H   | W   | C | S | I | V | Y | Y | E | L | N | N | R | V | G | E | A | F | H | A | S | S | T | S | L | D | G | F | T | D | P | S | N | N | K | N | R | F | C | L | G | L |   |
| Smad1 X.tropicalis   | D   | V   | Q  | A  | A   | E   | P  | K  | H   | W   | C | S | I | V | Y | Y | E | L | N | N | R | V | G | E | A | F | H | A | S | S | T | S | L | D | G | F | T | D | P | S | N | N | R | N | R | F | C | L | G | L |   |
| Smad1 G.gallus       | D   | V   | Q  | A  | A   | E   | P  | K  | H   | W   | C | S | I | V | Y | Y | E | L | N | N | R | V | G | E | A | F | H | A | S | S | T | S | L | D | G | F | T | D | P | S | N | N | K | N | R | F | C | L | G | L |   |
| Smad1 D.erio         | D   | V   | H  | P  | A   | Q   | P  | K  | H   | W   | C | S | I | V | Y | Y | E | L | N | N | R | V | G | E | A | F | L | A | S | S | T | S | L | D | G | F | T | D | P | S | N | N | R | N | R | F | C | L | G | L |   |
| Smad4 S.scrofa       | D   | V   | Q  | A  | A   | E   | P  | K  | H   | W   | C | S | I | V | Y | Y | E | L | N | N | R | V | G | E | A | F | H | A | S | S | T | S | L | D | G | F | T | D | P | S | N | N | K | N | R | F | C | L | G | L |   |
| Smad5 H.sapiens      | D   | V   | Q  | P  | A   | E   | P  | K  | H   | W   | C | S | I | V | Y | Y | E | L | N | N | R | V | G | E | A | F | H | A | S | S | T | S | L | D | G | F | T | D | P | S | N | N | K | S | R | F | C | L | G | L |   |
| Smad5 M.musculus     | D   | V   | Q  | P  | A   | E   | P  | K  | H   | W   | C | S | I | V | Y | Y | E | L | N | N | R | V | G | E | A | F | H | A | S | S | T | S | L | D | G | F | T | D | P | S | N | N | K | S | R | F | C | L | G | L |   |
| Smad5 R.norvegicus   | D   | V   | Q  | P  | A   | E   | P  | K  | H   | W   | C | S | I | V | Y | Y | E | L | N | N | R | V | G | E | A | F | H | A | S | S | T | S | L | D | G | F | T | D | P | A | N | N | K | S | R | F | C | L | G | L |   |
| Smad5 G.gallus       | D   | V   | Q  | P  | A   | E   | P  | K  | H   | W   | C | S | I | V | Y | Y | E | L | N | N | R | V | G | E | A | F | H | A | S | S | T | S | L | D | G | F | T | D | P | S | N | N | K | N | R | F | C | L | G | L |   |
| Smad5 D.erio         | D   | V   | Q  | P  | E   | Q   | P  | S  | H   | W   | C | S | I | V | Y | Y | E | L | N | N | R | V | G | E | A | Y | H | A | S | S | T | S | L | D | G | F | T | D | P | S | N | N | K | N | R | F | C | L | G | L |   |
| Smad8 M.musculus     | D   | F   | R  | P  | C   | E   | P  | Q  | H   | W   | C | S | V | A | Y | Y | E | L | N | N | R | V | G | E | T | F | Q | A | S | S | R | S | L | D | G | F | T | D | P | S | N | N | R | N | R | F | C | L | G | L |   |
| Smad8 R.norvegicus   | D   | F   | R  | P  | C   | E   | P  | L  | H   | W   | C | S | V | A | Y | Y | E | L | N | N | R | V | G | E | T | F | Q | A | S | S | R | S | L | D | G | F | T | D | P | S | N | N | R | N | R | F | C | L | G | L |   |
| Smad8 G.gallus       | N   | F   | R  | P  | C   | E   | P  | Q  | H   | W   | C | S | V | A | Y | Y | E | L | N | N | R | V | G | E | T | F | Q | A | S | S | R | S | L | D | G | F | T | D | P | S | N | N | K | N | R | F | C | L | G | L |   |

AR

BR



# 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%   |

# Finding Low-harmony sites in Smad-MH2

| Pos. | Sec.str. | SH   | AR  | BR   | Interaction   |
|------|----------|------|-----|------|---------------|
| L263 | B1'      | 0    | La  | Vfm  | SARA          |
| T267 | B1'      | 0    | Tm  | Acen | SARA          |
| S269 | loop     | 0    | CSh | Ea   | ?? (putative) |
| A272 | loop     | 0    | A   | Kals | ?? (putative) |
| F273 | loop     | 0    | F   | Hy   | ?? (putative) |
| Q284 | B2       | 0    | Qt  | N    | TβR-I         |
| Q294 | loop     | 0.16 | Q   | Sq   | c-Ski/SnoN    |
| P295 | B3       | 0    | P   | Trl  | c-Ski/SnoN    |
| L297 | B3       | 0.11 | LMi | Vi   | c-Ski/SnoN    |
| T298 | B3       | 0    | T   | Li   | c-Ski/SnoN    |
| S308 | L1       | 0    | Sa  | N    | c-Ski/SnoN    |
| –    | L1       | 0    | –   | Nsd  | c-Ski/SnoN    |
| E309 | L1       | 0    | E   | Krs  | c-Ski/SnoN    |
| A323 | H1       | 0    | Ae  | S    | ALK1/2        |
| V325 | H1       | 0    | V   | I    | ?? (putative) |
| M327 | H1       | 0    | LMq | N    | ALK1/2        |
| R334 | Loop     | 0.18 | Rk  | K    | ?? (putative) |
| R337 | B5       | 0    | R   | H    | ?? (putative) |

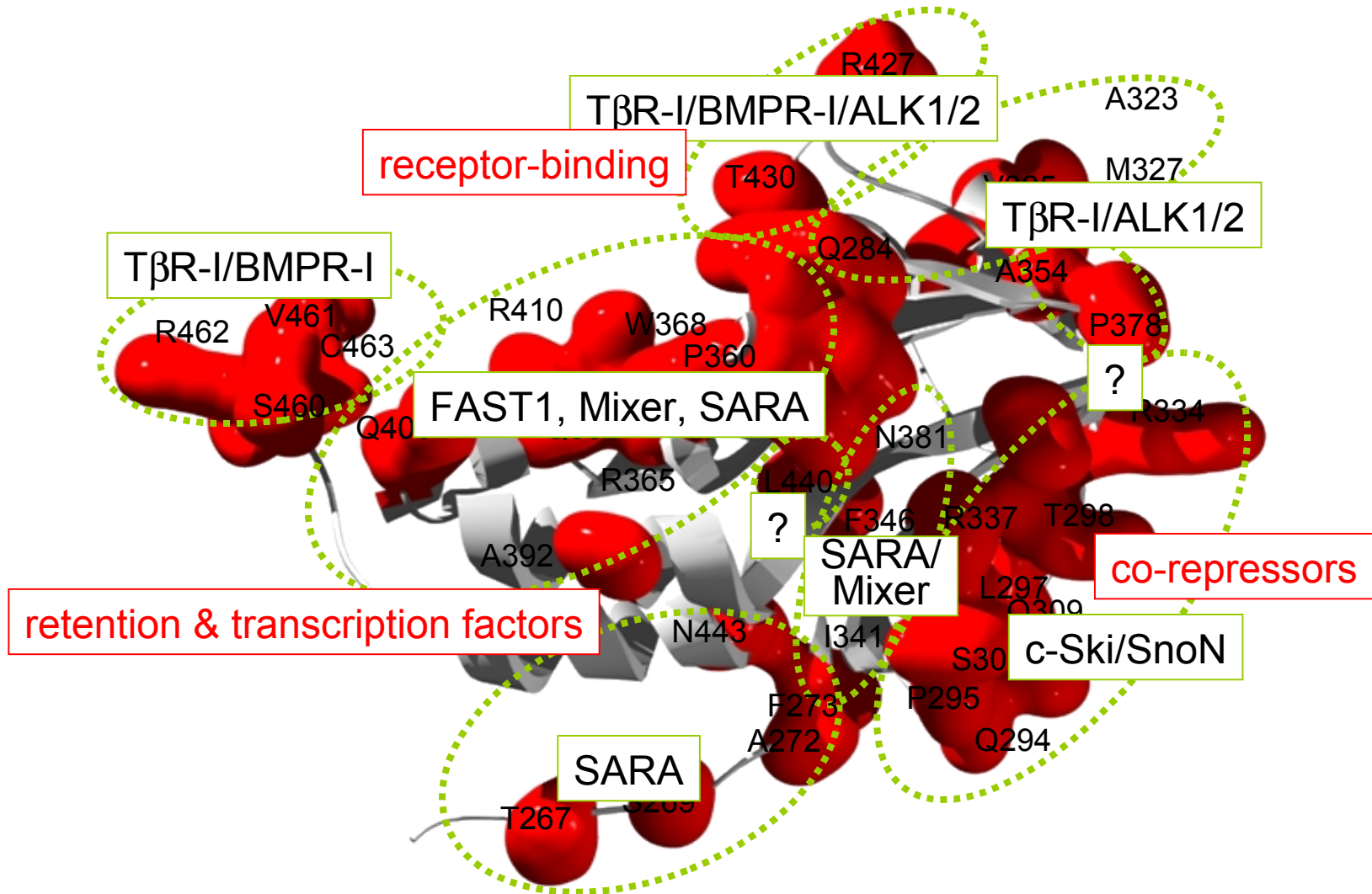
| Method           | Predict     | Specificity | Error |
|------------------|-------------|-------------|-------|
| AMAS             | 6           | 21%         | 3%    |
| TreeDet          | 21          | 52%         | 21%   |
| SDPpred          | 12          | 31%         | 10%   |
| Sequence Harmony | (SH=0) 32   | 79%         | 28%   |
|                  | (SH<0.2) 40 | 93%         | 33%   |

Pirovano, Feenstra & Heringa. “Sequence Comparison by Sequence Harmony Identifies Subtype Specific Functional Sites”, Nucleic Acids Res., in press (2006).

[www.few.vu.nl/~feenstra/articles/NAR\\_2006\\_Sequence\\_Harmony.pdf](http://www.few.vu.nl/~feenstra/articles/NAR_2006_Sequence_Harmony.pdf)



# Smad-MH2: Functional Clusters



# Conclusions Smad-MH2

- 40 Sites of Low Sequence Harmony in Smad-MH2
  - different between the AR (TGF- $\beta$ ) 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- $\beta$ /BMP specificity
  - confirm (or rebuke) putative functions?

# Sequence Harmony Webserver

Identification of Subtype Specific Functional Sites - SeaMonkey

http://www.ibi.vu.nl/programs/seqharmwww1-b/

Centre for Integrative Bioinformatics VU  
vrije Universiteit amsterdam

IBVU - http://ibi.vu.nl/

Main page Research Teaching **Programs** Members Publications Vacancies Related sites

## Sequence Harmony

Sequence Harmony is an entropy-based method, which accurately detects subfamily specific functional sites from alignment. The algorithm implements a new compositional differences between subfamilies on an intuitive scale.

Information on the [input format](#)

[Method Background](#)

Pirovano WA<sup>1</sup>, Feenstra KA<sup>1</sup> and Heringa J. (2006) Sequence Comparison by Sequence Harmony Identifies Subtype Specific Functional Sites. Nucl. Acids Res., in press.  
<sup>1</sup> joint first authors

Paste in a multiple alignment in **fasta** format:

Or Upload a multiple alignment in **fasta** format:

(you can use the [small example](#) for a test run)

Or  to select from input sequence names.

# Sequence Harmony Webserver: Groups

Identification of Subtype Specific Functional Sites - SeaMonkey

File Edit View Go Bookmarks Tools Window Help

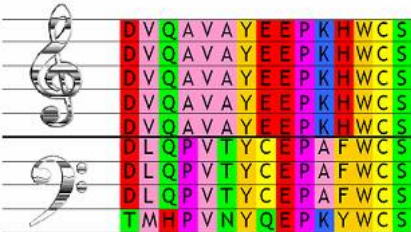
http://www.ibi.vu.nl/programs/seqharmwww1-b/

Centre for Integrative Bioinformatics VU  
vrije Universiteit amsterdam

IBIVU - http://ibi.vu.cs.vu.nl/

Main page Research Teaching **Programs** Members Publications Vacancies Related sites

## Sequence Harmony



Sequence Harmony is an entropy-based method, which accurately detects subfamily specific functional sites from a multiple sequence alignment. The algorithm implements a new formula, able to score compositional differences between subfamilies in a simple manner on an intuitive scale.

Information on the [Input format](#)

[Method Background](#)

Pirovano WA<sup>1</sup>, Feenstra KA<sup>1</sup> and Herin. Comparison by Sequence Harmony Ide. Functional Sites. Nucl. Acids Res., in press. .  
<sup>1</sup> joint first authors

Paste in a multiple alignment in **fasta** format:

```
>2_2967646
DLQPTVYSEPAFWCSIAYYELNQRVGTEFHASQPSLTVDGFTDPSNS-ERFCLGLL
GRGVRLLYYIGGEVFAECLSDSAIFVQSPNCNQRYGWHPATVCKIPPGCNLKFNNQ
GFEAVYQLTRMCTIRMSFVKGWGAEYRRQTVTSTPCWIELHLNGLQWLDKVLTMQ
>2_45331050
DLQPTVYSEPAFWCSIAYYELNQRVGTEFHASQPSLTVDGFTDPSNS-ERFCLGLL
GRGVRLLYYIGGEVFAECLSDSAIFVQSPNCNQRYGWHPATVCKIPPGCNLKFNNQ
GFEAVYQLTRMCTIRMSFVKGWGAEYRRQTVTSTPCWIELHLNGLQWLDKVLTMQ
>2_5360217
DLQPTVYSEPAFWCSIAYYELNQRVGTEFHASQPSLTVDGFTDPSNS-ERFCLGLL
GRGVRLLYYIGGEVFAECLSDSAIFVQSPNCNQRYGWHPATVCKIPPGCNLKFNNQ
GFEAVYQLTRMCTIRMSFVKGWGAEYRRQTVTSTPCWIELHLNGLQWLDKVLTMQ
```

Or Upload a multiple alignment in **fasta** format:

Select identifier of the first sequence of the second subgroup:

1\_7160686 Or Make List again.

2\_2967646

3\_45331050

2\_6288775  
2\_58047721  
2\_4009524  
2\_40254710  
2\_1KHX  
3\_13992583  
3\_11875329  
3\_45331052  
3\_18418623  
3\_2564493  
3\_6981174  
3\_45383213  
3\_47523074  
1\_7160686  
1\_61967924  
1\_11464653  
1\_6273783

CENTRE FOR INTEGRATIVE  
BIOINFORMATICS VU

# Sequence Harmony Webserver: Reference

Identification of Subtype Specific Functional Sites - SeaMonkey

File Edit View Go Bookmarks Tools Window Help

http://www.ibi.vu.nl/programs/seqharmwww1-b/

Sequence Harmony is an entropy-based method, which accurately detects subfamily specific functional sites from a multiple sequence alignment. The algorithm implements a new formula, able to score compositional differences between subfamilies in a simple manner on an intuitive scale.

Information on the Input format

Method Background

Pirovano WA<sup>1</sup>, Feenstra KA<sup>1</sup> and Heringa J. (2006) Sequence Comparison by Sequence Harmony Identifies Subtype Specific Functional Sites. *Nucl. Acids Res.*, in press.

<sup>1</sup> joint first authors

Sequence Harmony (basic)

Or Upload a multiple alignment in *fasta* format:

Browse...

(you can use the [small example](#) for a test-ride)  
(Your file selection was: Smad\_23158+1KHX\_MH2\_all\_Muscle.fa)

Select identifier of the first sequence of the second subgroup:  
1\_716086 Or Make List again.

Submit Clear Input

Advanced Features

**Reference Sequence (optional):**  
Identifier: 2\_1KHX and starting position 261  
(starting position indicates the number of first residue in the structure, this will be positive. If your structure contains only part of the sequence, this may be negative.)

**Reference Structure (optional):**  
PDB identifier OR PDB file  
Browse...

**Note:** if the sequence and numbering of your reference sequence do not correspond exactly to that of the reference sequence, you can correct that using the 'starting position' field. If you make a mistake there, the 'B-factor coloring' produced in your structure will not be meaningful.

Submit Clear Input

The Sequence Harmony web server was developed by K. Anton Feenstra and Walter Pirovano

(c) IBIVU 2006. If you are experiencing problems with the site, please contact the [webmaster](#).

Done



# Sequence Harmony Webserver: Structure

Identification of Subtype Specific Functional Sites - SeaMonkey

File Edit View Go Bookmarks Tools Window Help

http://www.ibi.vu.nl/programs/seqharmwww1-b/

DLQPVITYCEPAFWCS  
DLQPVITYCEPAFWCS  
IMHPVNYQEPKYWCS

Sequence Harmony is an entropy-based method, which accurately detects subfamily specific functional sites from a multiple sequence alignment. The algorithm implements a new formula, able to score compositional differences between subfamilies in a simple manner on an intuitive scale.

Information on the Input format

Method Background

Pirovano WA<sup>1</sup>, Feenstra KA<sup>1</sup> and Heringa J. (2006) Sequence Comparison by Sequence Harmony Identifies Subtype Specific Functional Sites. Nucleic Acids Res., in press.

<sup>1</sup> joint first authors

Sequence Harmony (basic)

Or Upload a multiple alignment in *fasta* format:

Browse...

(you can use the [small example](#) for a test-ride)  
(Your file selection was: Smad\_23158+1KHX\_MH2\_all\_Muscle.fa)

Select identifier of the first sequence of the second subgroup:  
1\_7160686 Or Make List again.

Submit Clear Input

Advanced Features

Sequence Harmony cutoff value: 0.2  
(values between 0-1 make sense)

Reference Sequence (optional):

PDB identifier 1KHX OR PDB file

Browse...

Note: if the sequence and numbering of your reference structure do not correspond exactly to that of the reference sequence, you can correct that in the 'field'. If you make a mistake there, the used in your structure will not be meaningful.

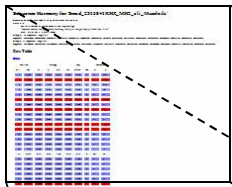
Submit

The Sequence Harmony web server was developed by K. Anton Feenstra and Walter Pirovano

(c) IBIVU 2006. If you are experiencing problems with the site, please contact the [webmaster](#).



# SMAD Sequence Harmony: Raw Table



Sequence Harmony for 'Smad\_23158+1KHX\_MH2\_ali\_Muscle.fa' - SeaMonkey

File Edit View Go Bookmarks Tools Window Help

http://www.ibi.vu.nl/jobs/4b209033e4a496d7f48f5a71a252b3e/raw\_table.html

Go Search

Sequence Harmony for 'Smad\_23158+1KHX\_MH2\_ali\_Muscle.fa'

Generated on Tue 21 Nov 2006 11:45 by CalculateSH version 0.1b

Please cite:

Walter Pirovano, K. Anton Feenstra and Jaap Heringa

"Sequence Comparison by Sequence Harmony Identifies Subtype Specific Functional Sites"

Nucl. Acids Res., in press (2006)

Group A: 16 sequences, length 211

Sequences: 2\_2967646, 2\_45331050, 2\_5360217, 2\_6288775, 2\_58047721, 2\_4009524, 2\_40254710, 2\_1KHX, 3\_13992583, 3\_11875329, 3\_45331052, 3\_18418623, 3\_2564493, 3\_6981174, 3\_45383213, 3\_47523074

Group B: 17 sequences, length 211

Sequences: 1\_7160686, 1\_61967924, 1\_11464653, 1\_6273783, 1\_11907945, 1\_29725652, 1\_1654323, 1\_3192871, 1\_55926152, 5\_61967926, 5\_2360958, 5\_5706366, 5\_3982649, 5\_6288777, 8\_61967928, 8\_2689629, 8\_22532988

Raw Table

Results

| Position |      | Entropy |      |      | Seq. |          | Consensus |    |
|----------|------|---------|------|------|------|----------|-----------|----|
| Ali      | Ref  | A       | B    | AB   | rel. | Harm.Rnk | A         | B  |
| 1        | D261 | 0.00    | 0.64 | 0.39 | 0.06 | 0.94     | 0         | D  |
| 2        | L262 | 0.34    | 0.98 | 1.67 | 1.00 | 0.00     | 1         | La |
| 3        | Q263 | 0.34    | 1.16 | 0.95 | 0.19 | 0.81     | 0         | Qa |
| 4        | P264 | 0.00    | 0.98 | 0.75 | 0.24 | 0.76     | 0         | P  |
| 5        | V265 | 0.00    | 0.00 | 0.00 | 0.00 | 1.00     | 0         | V  |
| 6        | T266 | 0.34    | 1.28 | 1.82 | 1.00 | 0.00     | 2         | Tm |
| 7        | Y267 | 0.00    | 0.00 | 0.00 | 0.00 | 1.00     | 0         | Y  |
| 8        | S268 | 1.27    | 0.67 | 1.96 | 1.00 | 0.00     | 2         | CS |
| 9        | E269 | 0.00    | 0.00 | 0.00 | 0.00 | 1.00     | 0         | E  |
| 10       | P270 | 0.34    | 0.00 | 0.20 | 0.03 | 0.97     | 0         | Ps |
| 11       | A271 | 0.00    | 1.14 | 1.59 | 1.00 | 0.00     | 1         | A  |
| 12       | F272 | 0.00    | 0.32 | 1.17 | 1.00 | 0.00     | 2         | F  |
| 13       | W273 | 0.00    | 0.00 | 0.00 | 0.00 | 1.00     | 0         | W  |
| 14       | C274 | 0.00    | 0.00 | 0.00 | 0.00 | 1.00     | 0         | C  |
| 15       | S275 | 0.00    | 0.00 | 0.00 | 0.00 | 1.00     | 0         | S  |
| 16       | I276 | 0.00    | 0.67 | 0.44 | 0.09 | 0.91     | 0         | I  |
| 17       | A277 | 0.99    | 0.67 | 1.56 | 0.73 | 0.27     | 0         | SA |
| 18       | Y278 | 0.00    | 0.00 | 0.00 | 0.00 | 1.00     | 0         | Y  |
| 19       | Y279 | 0.00    | 0.00 | 0.00 | 0.00 | 1.00     | 0         | Y  |
| 20       | E280 | 0.00    | 0.00 | 0.00 | 0.00 | 1.00     | 0         | E  |

Done

# SMAD Sequence Harmony: Results

| Position |      | Entropy |      |      | Seq. |          | Consensus |    |
|----------|------|---------|------|------|------|----------|-----------|----|
| Ali      | Ref  | A       | B    | AB   | rel. | Harm.Rnk | A         | B  |
| 106      | Y365 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 4         | Y  |
| 108      | W367 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 4         | W  |
| 207      | C462 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 4         | C  |
| 37       | T297 | 0.00    | 0.32 | 1.17 | 1.00 | 0.00     | 4         | T  |
| 105      | R364 | 0.00    | 0.67 | 1.35 | 1.00 | 0.00     | 4         | R  |
| 104      | Q363 | 0.00    | 0.79 | 1.40 | 1.00 | 0.00     | 4         | Q  |
| 34       | P294 | 0.00    | 0.98 | 1.50 | 1.00 | 0.00     | 4         | P  |
| 205      | V460 | 1.42    | 0.00 | 1.69 | 1.00 | 0.00     | 4         | IV |
| 65       | V324 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 3         | V  |
| 63       | A322 | 0.34    | 0.00 | 1.16 | 1.00 | 0.00     | 3         | Ae |
| 47       | S307 | 0.34    | 0.00 | 1.16 | 1.00 | 0.00     | 3         | Sa |
| 49       | E308 | 0.00    | 0.98 | 1.50 | 1.00 | 0.00     | 3         | E  |
| 48       | -    | 0.00    | 1.09 | 1.56 | 1.00 | 0.00     | 3         | -  |
| 67       | M326 | 1.27    | 0.00 | 1.62 | 1.00 | 0.00     | 3         | LM |
| 12       | F272 | 0.00    | 0.32 | 1.17 | 1.00 | 0.00     | 2         | F  |
| 11       | A271 | 0.00    | 1.14 | 1.59 | 1.00 | 0.00     | 2         | A  |
| 6        | T266 | 0.34    | 1.28 | 1.82 | 1.00 | 0.00     | 2         | Tm |
| 8        | S268 | 1.27    | 0.67 | 1.96 | 1.00 | 0.00     | 2         | CS |
| 86       | F345 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 1         | F  |
| 100      | P359 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 1         | P  |
| 121      | N380 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 1         | N  |
| 144      | Q399 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 1         | Q  |
| 154      | R409 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 1         | R  |
| 171      | R426 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 1         | R  |
| 174      | T429 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 1         | T  |
| 77       | R336 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 1         | R  |
| 81       | I340 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 1         | I  |
| 151      | Q406 | 0.34    | 0.00 | 1.16 | 1.00 | 0.00     | 1         | Qr |
| 23       | Q283 | 0.34    | 0.00 | 1.16 | 1.00 | 0.00     | 1         | Qt |
| 184      | L439 | 0.00    | 0.32 | 1.17 | 1.00 | 0.00     | 1         | L  |
| 136      | A391 | 0.00    | 0.64 | 1.33 | 1.00 | 0.00     | 1         | A  |
| 2        | L262 | 0.34    | 0.98 | 1.67 | 1.00 | 0.00     | 1         | La |
| 204      | S459 | 0.67    | 1.17 | 1.86 | 0.94 | 0.06     | 4         | Sn |
| 36       | L296 | 1.27    | 0.79 | 1.91 | 0.89 | 0.11     | 4         | LM |
| 33       | Q293 | 0.00    | 0.32 | 1.00 | 0.84 | 0.16     | 4         | Q  |
| 187      | N442 | 0.00    | 0.32 | 1.00 | 0.84 | 0.16     | 1         | N  |
| 118      | P377 | 0.00    | 0.32 | 1.00 | 0.84 | 0.16     | 1         | P  |
| 206      | R461 | 0.34    | 0.00 | 0.99 | 0.83 | 0.17     | 4         | Rp |
| 94       | A353 | 0.34    | 0.00 | 0.99 | 0.83 | 0.17     | 1         | As |
| 74       | R333 | 0.34    | 0.00 | 0.99 | 0.83 | 0.17     | 1         | Rk |
| Average  |      | 0.04    | 0.06 | 0.24 | 0.18 | 0.01     |           |    |

Sequence Harmony for 'Smad\_23158+1KHX\_MH2\_ali\_Muscle.fa' - SeaMonkey

Sequence Harmony for 'Smad\_23158+1KHX\_MH2\_ali\_Muscle.fa'

Generated on Tue 21 Nov 2006 11:49 by Calculate5H version 0.1b  
Please cite:  
Walter Pirovano, K. Anton Feenstra and Jaap Heringa  
*"Sequence Comparison by Sequence Harmony Identifies Subtype Specific Functional Sites"*  
Nucl. Acids Res., in press (2006)  
Group A: 16 sequences, length 211  
Sequences: 2\_2967646, 2\_45331050, 2\_5360217, 2\_6288775, 2\_58047721, 2\_4009524, 2\_40294710, 2\_1KHX, 3\_13992583, 3\_11875329, 3\_45331052, 3\_18418623, 3\_2564493, 3\_6981174, 3\_45383213, 3\_47523074  
Group B: 17 sequences, length 211  
Sequences: 1\_7160686, 1\_61967924, 1\_11464653, 1\_6273783, 1\_11907945, 1\_29725652, 1\_1654323, 1\_3192871, 1\_55926152, 5\_61967926, 5\_2360958, 5\_5706366, 5\_3982649, 5\_6288777, 8\_61967928, 8\_2689629, 8\_22532988  
testing

Results

Raw Table

| Position |      | Entropy |      |      | Seq. |          | Consensus |    |
|----------|------|---------|------|------|------|----------|-----------|----|
| Ali      | Ref  | A       | B    | AB   | rel. | Harm.Rnk | A         | B  |
| 106      | Y365 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 4         | Y  |
| 108      | W367 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 4         | W  |
| 207      | C462 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 4         | C  |
| 37       | T297 | 0.00    | 0.32 | 1.17 | 1.00 | 0.00     | 4         | T  |
| 105      | R364 | 0.00    | 0.67 | 1.35 | 1.00 | 0.00     | 4         | R  |
| 104      | Q363 | 0.00    | 0.79 | 1.40 | 1.00 | 0.00     | 4         | Q  |
| 34       | P294 | 0.00    | 0.98 | 1.50 | 1.00 | 0.00     | 4         | P  |
| 205      | V460 | 1.42    | 0.00 | 1.69 | 1.00 | 0.00     | 4         | IV |
| 65       | V324 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 3         | V  |
| 63       | A322 | 0.34    | 0.00 | 1.16 | 1.00 | 0.00     | 3         | Ae |
| 47       | S307 | 0.34    | 0.00 | 1.16 | 1.00 | 0.00     | 3         | Sa |
| 49       | E308 | 0.00    | 0.98 | 1.50 | 1.00 | 0.00     | 3         | E  |
| 48       | -    | 0.00    | 1.09 | 1.56 | 1.00 | 0.00     | 3         | -  |
| 67       | M326 | 1.27    | 0.00 | 1.62 | 1.00 | 0.00     | 3         | LM |
| 12       | F272 | 0.00    | 0.32 | 1.17 | 1.00 | 0.00     | 2         | F  |
| 11       | A271 | 0.00    | 1.14 | 1.59 | 1.00 | 0.00     | 2         | A  |
| 6        | T266 | 0.34    | 1.28 | 1.82 | 1.00 | 0.00     | 2         | Tm |
| 8        | S268 | 1.27    | 0.67 | 1.96 | 1.00 | 0.00     | 2         | CS |
| 86       | F345 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 1         | F  |
| 100      | P359 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 1         | P  |
| 121      | N380 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 1         | N  |
| 144      | Q399 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 1         | Q  |
| 154      | R409 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 1         | R  |
| 171      | R426 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 1         | R  |
| 174      | T429 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 1         | T  |
| 77       | R336 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 1         | R  |
| 81       | I340 | 0.00    | 0.00 | 1.00 | 1.00 | 0.00     | 1         | I  |
| 151      | Q406 | 0.34    | 0.00 | 1.16 | 1.00 | 0.00     | 1         | Qr |
| 23       | Q283 | 0.34    | 0.00 | 1.16 | 1.00 | 0.00     | 1         | Qt |
| 184      | L439 | 0.00    | 0.32 | 1.17 | 1.00 | 0.00     | 1         | L  |
| 136      | A391 | 0.00    | 0.64 | 1.33 | 1.00 | 0.00     | 1         | A  |
| 2        | L262 | 0.34    | 0.98 | 1.67 | 1.00 | 0.00     | 1         | La |
| 204      | S459 | 0.67    | 1.17 | 1.86 | 0.94 | 0.06     | 4         | Sn |
| 36       | L296 | 1.27    | 0.79 | 1.91 | 0.89 | 0.11     | 4         | LM |
| 33       | Q293 | 0.00    | 0.32 | 1.00 | 0.84 | 0.16     | 4         | Q  |
| 187      | N442 | 0.00    | 0.32 | 1.00 | 0.84 | 0.16     | 1         | N  |
| 118      | P377 | 0.00    | 0.32 | 1.00 | 0.84 | 0.16     | 1         | P  |
| 206      | R461 | 0.34    | 0.00 | 0.99 | 0.83 | 0.17     | 4         | Rp |
| 94       | A353 | 0.34    | 0.00 | 0.99 | 0.83 | 0.17     | 1         | As |
| 74       | R333 | 0.34    | 0.00 | 0.99 | 0.83 | 0.17     | 1         | Rk |

# SMAD Sequence Harmony: Structure

