Poster H-36

Sequence Comparison by Sequence Harmony Identifies Sub-type Specific Sites for Smad Receptor Binding

Authors:
Walter Pirovano (*IBIVU, Free University Amsterdam*)
K. Anton Feenstra (*IBIVU, Free University Amsterdam*)
Jaap Heringa (*IBIVU, Free University Amsterdam*)

Short Abstract: We introduce a new entropy-based method, called Sequence Harmony, to accurately detect subclass-specific positions based on compositional differences. Applying the method on the Smad-family of TFs yields 40 sites of which 26 have a known function. Structural considerations led to the assignment of a putative function of 11 more sites.

Long Abstract:
Multiple sequence alignments (MSAs) are often used to reveal functionally important residues within a protein family. In particular, they can be very useful for identification of key residues that determine functional differences between protein subclasses. Starting from a MSA of proteins of interest, the aim is to identify sites that are possibly conserved within a subclass, but certainly different between the groups. We will derive from Shannon's general information entropy as applied to sequences by Shenkin et al. an alternative similarity measure named Sequence Harmony (SH) for comparison of groups of sequences within a multiple sequence alignment. The Sequence Harmony has well-defined properties, is easily calculated, not dependent on the relative size of the groups, yields values on a convenient interval of [0...1] and intuitively corresponds to differences in the aminoacid compositions as observed in the alignment. In the transforming growth factor-β (TGF-β) signalling pathway the Smad family of transcription factors play a crucial role and are critical to determine the specificity between similar pathways. This complex signalling network is involved in regulation of many cellular processes like division and differentiation, motility, adhesion and programmed death. The TGF-β family includes several members: TGF-βs, the nodals, the activins and the bone morphogenetic proteins (BMPs). Over the last years a great number of experiments have yielded much information about specific receptor-Smad interactions and the role of numerous additional proteins involved in this process. On the other hand, there is still much to learn about the specific interactions of the Smads with other factors that determine the specificity of TGF-β and BMP–associated pathways. It has been shown that most of these interactions map to the conserved Mad Homology 2 (MH2) domain of the Smad proteins. Using Sequence Harmony, we will systematically and specifically identify all sites associated with the TGF-β vs. BMP subclass-specificity of Smads MH2 domain. Out of the 211 residues in the MH2 sequence alignment, only 40 have a low Sequence Harmony (SH ≤0.2) of which 26 have a known function (65%). For the 32 non-harmonious sites (SH zero), 22 have a known function (69%). Of the 171 remaining high-harmony sites, on the other hand, to the best of our knowledge only very few specific sites have been identified as important for specific receptor binding. Interestingly, 19 of the 40 low harmony sites are not completely conserved within either group, but still are completely different between the groups. Here, the strength of the

sequence harmony method becomes apparent. The Sequence Harmony data was also projected onto a crystal structure of a Smad MH2 domain. From this we can identify a limited number of clustered regions high in low-harmony sites. Three of these clusters are associated with receptor binding. The second-largest cluster is associated with c-Ski/SnoN interactions, which are responsible for inhibition of TGFβ-activated transcription. Three other clusters are associated with the Smad-interacting transcription factors FAST1, Mixer or with the cytosolic retention factor SARA. Taking the spatial clusters observed as a guideline, we can assign putative functions to 11 out of the 14 residues of unknown function. Four unknowns can be assigned a putative function in FAST1, Mixer and/or SARA binding. Four other unknowns have a putative function in SARA binding. Two unknowns have a putative function in co-repressor (c-Ski/SnoN) binding. One unknown has a putative function in receptor-binding. Concluding, sites of low Sequence Harmony correspond very specifically to functionally relevant sites in the Smad-MH2 domain, with a very sharp separation between conserved positions (which are the majority) and those that show a clear difference between the TGF-b and the BMP-binding sub-types. The scale of the sequence harmony can intuitively be interpreted as sites that are more or less likely to be of functional importance. We have identified 14 sites of low sequence harmony, which have not been experimentally verified. We would hereby suggest these as promising candidates for further elucidation of their function in determining the specificity of the TGF-b and BMP-associated signalling pathways. Specifically, it would be very interesting to confirm (or rebuke) the putative functional roles we assigned them based on their spatial clustering with low harmony sites of known function.