Sequence comparison by sequence harmony identifies subtype-specific functional sites

Walter Pirovano, K. Anton Feenstra and Jaap Heringa*

Centre for Integrative Bioinformatics VU (IBIVU), Vrije Universiteit, De Boelelaan 1081A, 1081 HV Amsterdam, The Netherlands

Received July 10, 2006; Revised October 12, 2006; Accepted October 13, 2006

ABSTRACT

Multiple sequence alignments are often used to reveal functionally important residues within a protein family. They can be particularly useful for the identification of key residues that determine functional differences between protein subfamilies. We present a new entropy-based method, Sequence Harmony (SH) that accurately detects subfamily-specific positions from a multiple sequence alignment. The SH algorithm implements a novel formula, able to score compositional differences between subfamilies, without imposing conservation, in a simple manner on an intuitive scale. We compare our method with the most important published methods, i.e. AMAS, TreeDet and SDP-pred, using three well-studied protein families: the receptor-binding domain (MH2) of the Smad family of transcription factors, the Ras-superfamily of small GTPases and the MIP-family of integral membrane transporters. We demonstrate that SH accurately selects known functional sites with higher coverage than the other methods for these test-cases. This shows that compositional differences between protein subfamilies provide sufficient basis for identification of functional sites. In addition, SH selects a number of sites of unknown function that could be interesting candidates for further experimental investigation.

INTRODUCTION

In the quest for knowledge about protein function, understanding differences between protein families is essential. It is therefore not surprising that a large number of methods have already been introduced for the positional comparison of amino acid compositions between different protein families or subtypes [1–9]; for a review see Whisstock and Lesk (10)]. Though these methods have contributed greatly to the understanding of the relation between sequence and function (11), coverage of known sites of functional differences has been limited.

An early method called AMAS by Livingstone and Barton (1) analyses conservation patterns using a number of physicochemical properties of amino acids. The method assigns sites to either of three classes: globally conserved, conserved in subfamilies, or not conserved, based on arbitrarily set conservation threshold values. This method was intended primarily to ‘allow the residue-specific similarities and differences in physicochemical properties between groups of sequences to be identified quickly’ (1).

Several methods have been based on evolutionary trace analysis (2) and rely on residue conservation within subfamilies to construct subfamily-specific consensus sequences. These sequences are then aligned to reveal the variation between the different subfamilies. Kuipers et al. (3) relaxed the requirement of intra-group conservation and allowed the selection of ‘residues that are conserved in one class of proteins with a certain function but are different in other classes’. Del Sol Mesa et al. (6) introduced a method called TreeDet that uses mutational behaviour analysis of so-called ‘tree-determinant’ residues. The method uses an internal algorithm for unsupervised grouping of the input sequences. It then selects sites that follow mutation patterns similar to that of the overall phylogeny. For each alignment position, this is measured using the correlation coefficient between the substitution matrix derived for the position considered and that for the whole protein. Residues selected tend to be conserved within each subfamily but different between them. The method aims to find ‘the most appropriate way of dividing a protein family into subfamilies in order to associate the tree-determinants with sites, which are likely to be responsible for functional differences between these subfamilies’.

The method SDP-pred by Kalinina et al. (7) selects residues that ‘are well conserved within specificity groups but differ between these groups’. The method is based on mutual information analysis and complex statistical treatment. It extends an earlier method (5) by using residue frequencies smoothed and weighted by the BLOSUM average substitution scores. Significance is estimated by Z-scores of...
expected mutual information obtained from column shuffling. Subsequently, an appropriate Z-score threshold for selection of high-ranking sites is determined using the Bernoulli estimator. Recently, Donald and Shakhnovich (8) added an automated procedure for functional grouping.

The above methods focus on sites that are conserved in one or both groups and subsequently select those sites that are different between these groups (10). Unfortunately, this scenario excludes sites that are not highly conserved within each of the groups. This may not seem a serious problem at first hand, but let us consider an example. Take a group A comprising a protein subfamily binding a certain molecule (e.g. a ligand or receptor; this group represents the ‘binders’) and group B comprising proteins that do not (the ‘non-binders’). Certainly, one can expect sites that are crucial for binding to be conserved in group A (the ‘binders’). However, for group B to avoid binding, the corresponding site would need to avoid the conserved residue of group A. It seems therefore imprudent to expect conservation throughout group B (the ‘non-binders’) as well (3). Moreover, if group A contains binders to different (but related) molecules (ligands or receptors), even the requirement of conservation in group A may not apply.

To address these restrictions here, we introduce an alternative similarity measure for comparing groups of sequences within a multiple sequence alignment, which we name Sequence Harmony (SH). The SH measure is derived from Shannon’s general information entropy (12) as applied to biomolecular sequences by Shenkin et al. (13). We will show that SH has well-defined properties and is easily calculated. The values obtained fall within a convenient fixed interval and correspond intuitively to differences in the amino acid compositions as observed in the alignment. Moreover, no additional residue substitution matrices, calculations of mutual information or computations of statistical significance are needed (5,7).

We evaluate the SH method using four benchmark sets comprising experimental data on specificity-switching residues and compare its performance against three other state-of-the-art methods. From this, we obtain new insights about the principles that govern the accurate detection of functional sites. We will show that SH is able to identify sites associated with subfamily specificity systematically and with relatively few errors.

**THEORY**

**Comparing sequences by relative entropy**

Relative entropy (rE) is commonly used in sequence comparison to quantify the degree of conservation (14,15). It is derived from Shannon’s general information entropy (12) as applied to biological sequences by Shenkin et al. (13):

$$rE_{i}^{AB} = \sum_{x} p_{i,x}^{A} \log \frac{p_{i,x}^{A}}{p_{i,x}^{B}}$$  \hspace{1cm} (1)

where $p_{i,x}^{A}$ and $p_{i,x}^{B}$ are the observed probabilities of amino acid type $x$ at a position $i$ in the alignment of groups A and B, respectively. Relative entropy measures the difference in information content between both distributions of amino acid types. Interestingly, for sites to be maximally different between the two groups, amino acid types in one group should be absent in the other or vice versa. This leads to a degenerate result (singularity) whenever the entropy function of Equation 1 is used. Inclusion of so-called ‘pseudo-counts’ (16) solves the degeneracy but not the unbounded, asymptotic behaviour of Equation 1. Also taking, for instance, the relative entropy with respect to both groups, as $rE_{i}^{A/B} = \sum_{x} p_{i,x}^{A} \log (p_{i,x}^{A}/p_{i,x}^{B})$, does not solve the unbounded behaviour. Moreover, the upper limit depends on the ratio of the number of sequences in both groups.

**Calculating differences by SH**

We address the degeneracy of Equation 1 by defining SH as

$$SH_{i}^{AB} = \sum_{x} p_{i,x}^{A} \log \frac{p_{i,x}^{A}}{p_{i,x}^{A} + p_{i,x}^{B}}$$  \hspace{1cm} (2)

This can be viewed as the relative entropy of group A relative to the sum of the probabilities of both groups ($p^{A} + p^{B}$). As a consequence of the separate weighting of both groups, we eliminate the dependence on relative group sizes. Since, in general, $SH_{i}^{A/B} \neq SH_{i}^{B/A}$, we remedy this non-commutativity by taking the average. Using Equation 2 this leads to:

$$SH_{i} = \frac{1}{2} (SH_{i}^{AB} + SH_{i}^{BA})$$

$$= \frac{1}{2} \left( \sum_{x} p_{i,x}^{A} \log p_{i,x}^{A} + \sum_{x} p_{i,x}^{B} \log p_{i,x}^{B} \right)$$

$$- \sum_{x} (p_{i,x}^{A} + p_{i,x}^{B}) \log (p_{i,x}^{A} + p_{i,x}^{B})$$  \hspace{1cm} (3)

Using Shannon’s information entropy:

$$S_{i} = -\sum_{x} p_{i,x} \log p_{i,x}$$  \hspace{1cm} (4)

and writing

$$S_{i}^{A+B} = -\sum_{x} (p_{i,x}^{A} + p_{i,x}^{B}) \log (p_{i,x}^{A} + p_{i,x}^{B})$$  \hspace{1cm} (5)

as Shannon’s entropy weighting groups A and B equally, we can rewrite Equation 3 as

$$SH_{i} = S_{i}^{A+B} - \frac{1}{2} (S_{i}^{A} + S_{i}^{B}).$$  \hspace{1cm} (6)

In other words, SH can also be conveniently expressed as a simple linear combination of entropies.

The SH function juxtaposes relative entropy since it becomes zero for maximally different sites and one for sites with identical distributions. We therefore coined the phrase Sequence Harmony (SH) as it indicates to what extent amino acid compositions between two groups of sequences are in harmony.

To illustrate how SH works in practice, we present a hypothetical alignment in Table 1. Positions 1, 2 and 3 all have a SH value of zero. In fact, the formula becomes zero any time the amino acid composition in one group is non-overlapping with that in the other group, regardless of conservation. At position 4, the amino acid of group A
(Ala) also occurs once in group B, yielding the lowest possible non-zero score for this number of sequences. For positions 5 and 6, there is an increasing overlap and hence growing SH values. Position 8 is conserved overall in the whole family and is therefore maximally harmonious with an SH value of one. Note that unconserved sites have SH = 1 whenever the two groups have identical compositions. This is illustrated in position 7, where equal proportions of R and K are shown.

The SH measure is implemented in a simple online server for calculating SH from an alignment. It can be accessed at www.ibi.vu.nl/programs/seqharmwww.

### Ranking identically scoring SH sites

The example of Table 1 shows that sites with different compositions can have identical SH values, which therefore cannot be ranked. This is particularly important for sites with SH = 0, as these are potentially discriminative for function. To address this issue, we derived a simple but effective ranking for identically scoring SH sites based on the distribution of low SH values (SH \( \leq 0.2 \)) over the alignment positions. Groups of sequentially adjacent low-harmony sites can have one intermediate high-harmony site in between, i.e. the sequence distance is two or less. The ranking is first on increasing SH (low harmony first), and then on decreasing group size (larger sequential groups first). Finally, sites that have equal SH scores and group sizes are ranked on the total entropy of the sites in both subfamilies. This scenario is implemented in the SH method which we will label ‘SH’.

We further explored the performance of an even simpler ranking on just SH values and entropy of the site, i.e. without ranking on sequential groups. This approach we will refer to as ‘SH/E’.

### Benchmarking

We compared predictions by SH with those obtained from three other methods over several protein families. For this purpose, we have used the online server of AMAS (http://barton.ebi.ac.uk/servers/amas_server.html) (1), the ‘mutational behaviour’ method available from the TreeDet server, which we will refer to as TreeDet/MB (http://somosierra.cnb.uam.es/Servers/treedetv2/) (6) and the SDP-pred server (http://math.genebee.msu.ru/~psn/) (5,7). The recent method by Donald and Shakhnovich (8) (see Introduction) could not be evaluated in this study due to lack of an online server.

The methods were evaluated over a number of test-sets (see below) using receiver–operator characteristics (ROC) plots of coverage (% recovered known functional sites) versus error (% wrong predictions). These were constructed from ranked lists of sites for SH and for SDP-pred. For AMAS, the conservation threshold was varied from 0 to 10 (see Introduction). For TreeDet/MB, the cut-off value was varied from 0 to 1.0. In both cases, small steps were taken so that the addition of single sites to the selection could be observed.

### DATASETS FOR VALIDATION

We have selected three relevant protein families, for which experimental evidence is available on subfamily-specific sites. These include families that have been used for construction, testing and/or validation of the other prediction methods and one family for which we have assembled an extensive dataset.

In many cases, the available experimental evidence concerns the exchange of a sequence segment, e.g. a loop or a helix. Positions within such segments that are conserved over both subfamilies were not regarded as subfamily specific. Although such positions are likely to be important for the function of the family, they are unable to explain functional differences between subfamilies at the residue level.

### TGF-β-associated transcription factors (Smad family)

The Smad family of transcription factors plays a crucial role in the transforming growth factor-β (TGF-β) signalling pathway. Smads are also critical for determining the specificity between alternative TGF-β pathways (see recent extensive reviews, see Feng and Derynck (17) and Massagué et al. (18)). This complex signalling network is involved in the regulation of many cellular processes such as division and differentiation, motility, adhesion and programmed cell death. The TGF-β family of growth factors induces Type-I transmembrane receptors to phosphorylate and activate the receptor-regulated Smads (R-Smads) (18). The R-Smads can be subdivided into two major groups: the AR-Smads, which are mainly induced by TGF-β-type receptors (TβR-I), and the BR-Smads, which are mainly induced by the BMP-type receptors (BMPR-I and ALK1/2). Subsequent associations among Smads are responsible for control of TGF-β target genes in the nucleus. It has been shown that most of the above interactions involve the so-called Mad Homology 2 (MH2) domain of the Smad proteins (17).

From an extensive literature search, we have identified 29 specific sites in the MH2 domain that are experimentally validated to be important for Smad specificity, as listed in Table 2.

### Small GTPases (RAS superfamily)

Members of the Ras superfamily of GTPases are implicated in the regulation of growth, survival, differentiation and...

---

**Table 1. Hypothetical alignment of two subfamilies**

<table>
<thead>
<tr>
<th>Alignment position</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Group A</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>seq1</td>
<td>R</td>
<td>E</td>
<td>L</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>K</td>
<td>K</td>
</tr>
<tr>
<td>seq2</td>
<td>R</td>
<td>E</td>
<td>L</td>
<td>A</td>
<td>A</td>
<td>F</td>
<td>K</td>
<td>K</td>
</tr>
<tr>
<td>seq3</td>
<td>R</td>
<td>E</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>Y</td>
<td>R</td>
<td>K</td>
</tr>
<tr>
<td>seq4</td>
<td>R</td>
<td>E</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>F</td>
<td>R</td>
<td>K</td>
</tr>
<tr>
<td><strong>Group B</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>seq1</td>
<td>H</td>
<td>N</td>
<td>V</td>
<td>A</td>
<td>A</td>
<td>Y</td>
<td>R</td>
<td>K</td>
</tr>
<tr>
<td>seq2</td>
<td>H</td>
<td>N</td>
<td>V</td>
<td>F</td>
<td>F</td>
<td>Y</td>
<td>R</td>
<td>K</td>
</tr>
<tr>
<td>seq3</td>
<td>H</td>
<td>N</td>
<td>F</td>
<td>F</td>
<td>A</td>
<td>F</td>
<td>R</td>
<td>K</td>
</tr>
<tr>
<td>seq4</td>
<td>H</td>
<td>S</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>Y</td>
<td>R</td>
<td>K</td>
</tr>
<tr>
<td>seq5</td>
<td>H</td>
<td>S</td>
<td>M</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>K</td>
<td>K</td>
</tr>
<tr>
<td>seq6</td>
<td>H</td>
<td>T</td>
<td>M</td>
<td>F</td>
<td>F</td>
<td>Y</td>
<td>K</td>
<td>K</td>
</tr>
<tr>
<td><strong>SH</strong></td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.35</td>
<td>0.54</td>
<td>0.79</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

For each position in the alignment the SH score between the subfamilies is calculated.
other processes in haematopoietic cells (19). They comprise six families. Experimental evidence for functional sites is available from the literature for the Ras versus Ral families (6,20). This set was used by Del Sol Mesa et al. (6) for the development of the TreeDet method. In addition, we include a test set of Rab5/6-specific sites obtained from (21,22).

Integral membrane transporters (MIP family)

Members of the MIP family are mainly involved in facilitating the transport of both water and small neutral solutes through the cellular membrane in all domains of life. There are about six MIP subfamilies, the two major being the aquaporins (AQPs) and the glycerol-uptake facilitators (GLPs) (23). The AQP and GLP subfamilies were used for the initial validation of the SDP-pred method (7). An arbitrary measure for functional significance of a site used by Kalinina et al. (7) was the proximity to the glycerol molecules that are bound inside the GLP pore channel in the crystal structure 1FX8 (24). Sites that were conserved in the training set of sequences were excluded. This scenario yielded a set of 23 putative functional sites closer than 5 Å to any of the three glycerol molecules (7).

Sequence retrieval and alignment

R-Smad protein sequences were collected using the NCBI query for sequence retrieval (www.ncbi.nih.gov). This resulted in 15 non-redundant sequences for AR-Smads and 17 for BR-Smads. All sequences were aligned using the PSI-Praline multiple sequence alignment online server (www.ibi.vu.nl/programs/pralinewww) (25,26). From the alignment obtained, the MH2 domain was selected for further analysis.

<table>
<thead>
<tr>
<th>Position</th>
<th>Sec. struc.</th>
<th>Consensus</th>
<th>SH</th>
<th>SH group</th>
<th>Other methods</th>
<th>Function</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Align</td>
<td>Smad2</td>
<td>AR</td>
<td>BR</td>
<td></td>
<td>AMAS</td>
<td>Tree Det</td>
<td>SDP pred</td>
</tr>
<tr>
<td>2</td>
<td>(L263)</td>
<td>B1’</td>
<td></td>
<td>Vfm</td>
<td>+</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>3</td>
<td>(Q264)</td>
<td>B1’</td>
<td></td>
<td>Qrh</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>6</td>
<td>T267</td>
<td>B1’</td>
<td></td>
<td>Acen</td>
<td>0</td>
<td>2</td>
<td>SARA (29)</td>
</tr>
<tr>
<td>8</td>
<td>S269</td>
<td>loop</td>
<td></td>
<td>CSmE</td>
<td>0</td>
<td>2</td>
<td>?(SARA)</td>
</tr>
<tr>
<td>11</td>
<td>A272</td>
<td>loop</td>
<td></td>
<td>Kqls</td>
<td>0</td>
<td>2</td>
<td>? (SARA)</td>
</tr>
<tr>
<td>12</td>
<td>F273</td>
<td>loop</td>
<td></td>
<td>Hbo</td>
<td>0</td>
<td>2</td>
<td>? (SARA)</td>
</tr>
<tr>
<td>23</td>
<td>Q284</td>
<td>B2</td>
<td></td>
<td>Qt</td>
<td>0</td>
<td>1</td>
<td>Tjb-1</td>
</tr>
<tr>
<td>33</td>
<td>Q294</td>
<td>loop</td>
<td></td>
<td>Sq</td>
<td>0.16</td>
<td>4</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>34</td>
<td>P295</td>
<td>B3</td>
<td></td>
<td>Trl</td>
<td>0</td>
<td>4</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>36</td>
<td>L297</td>
<td>B3</td>
<td></td>
<td>Vli</td>
<td>0.11</td>
<td>4</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>37</td>
<td>T298</td>
<td>B3</td>
<td></td>
<td>Li</td>
<td>0</td>
<td>4</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>47</td>
<td>S308</td>
<td>L1</td>
<td></td>
<td>N</td>
<td>0</td>
<td>3</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>48</td>
<td>L1</td>
<td>—</td>
<td></td>
<td>Nod</td>
<td>0</td>
<td>3</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>49</td>
<td>E309</td>
<td>L1</td>
<td></td>
<td>Krs</td>
<td>0</td>
<td>3</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>63</td>
<td>A323</td>
<td>H1</td>
<td></td>
<td>Ae</td>
<td>0</td>
<td>3</td>
<td>0.84</td>
</tr>
<tr>
<td>65</td>
<td>Y325</td>
<td>H1</td>
<td></td>
<td>Y</td>
<td>0</td>
<td>3</td>
<td>0.87</td>
</tr>
<tr>
<td>67</td>
<td>M327</td>
<td>H1</td>
<td></td>
<td>LMn</td>
<td>0</td>
<td>3</td>
<td>0.83</td>
</tr>
<tr>
<td>74</td>
<td>R334</td>
<td>loop</td>
<td></td>
<td>Rk</td>
<td>0.18</td>
<td>4</td>
<td>? (c-Ski/SnoN)</td>
</tr>
<tr>
<td>77</td>
<td>R337</td>
<td>B5</td>
<td></td>
<td>R</td>
<td>0</td>
<td>1</td>
<td>0.87</td>
</tr>
<tr>
<td>81</td>
<td>I341</td>
<td>B5</td>
<td></td>
<td>V</td>
<td>0</td>
<td>1</td>
<td>0.87</td>
</tr>
<tr>
<td>86</td>
<td>F346</td>
<td>B6</td>
<td></td>
<td>F</td>
<td>0</td>
<td>1</td>
<td>0.87</td>
</tr>
<tr>
<td>94</td>
<td>A354</td>
<td>B7</td>
<td></td>
<td>As</td>
<td>0.18</td>
<td>1</td>
<td>0.87</td>
</tr>
<tr>
<td>100</td>
<td>P360</td>
<td>H2</td>
<td></td>
<td>P</td>
<td>0</td>
<td>1</td>
<td>0.87</td>
</tr>
<tr>
<td>104</td>
<td>Q364</td>
<td>H2</td>
<td></td>
<td>Yf</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>105</td>
<td>R365</td>
<td>H2</td>
<td></td>
<td>Hq</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>106</td>
<td>Y366</td>
<td>H2</td>
<td></td>
<td>H</td>
<td>0</td>
<td>4</td>
<td>0.86</td>
</tr>
<tr>
<td>108</td>
<td>W368</td>
<td>loop</td>
<td></td>
<td>W</td>
<td>0</td>
<td>4</td>
<td>0.87</td>
</tr>
<tr>
<td>118</td>
<td>P378</td>
<td>loop</td>
<td></td>
<td>Sp</td>
<td>0.16</td>
<td>1</td>
<td>0.87</td>
</tr>
<tr>
<td>121</td>
<td>N381</td>
<td>B9</td>
<td></td>
<td>N</td>
<td>0</td>
<td>1</td>
<td>0.87</td>
</tr>
<tr>
<td>136</td>
<td>A392</td>
<td>H3</td>
<td></td>
<td>Aq</td>
<td>0</td>
<td>1</td>
<td>0.85</td>
</tr>
<tr>
<td>144</td>
<td>Q400</td>
<td>loop</td>
<td></td>
<td>QH</td>
<td>0</td>
<td>1</td>
<td>0.87</td>
</tr>
<tr>
<td>151</td>
<td>Q407</td>
<td>H4</td>
<td></td>
<td>Qr</td>
<td>0</td>
<td>1</td>
<td>0.83</td>
</tr>
<tr>
<td>154</td>
<td>R410</td>
<td>H4</td>
<td></td>
<td>R</td>
<td>0</td>
<td>1</td>
<td>0.87</td>
</tr>
<tr>
<td>171</td>
<td>R427</td>
<td>L3</td>
<td></td>
<td>H</td>
<td>0</td>
<td>1</td>
<td>0.87</td>
</tr>
<tr>
<td>174</td>
<td>T430</td>
<td>L3</td>
<td></td>
<td>T</td>
<td>0</td>
<td>1</td>
<td>0.87</td>
</tr>
<tr>
<td>184</td>
<td>L440</td>
<td>B11</td>
<td></td>
<td>L</td>
<td>0</td>
<td>1</td>
<td>0.84</td>
</tr>
<tr>
<td>187</td>
<td>N443</td>
<td>H5</td>
<td></td>
<td>N</td>
<td>0.16</td>
<td>1</td>
<td>0.83</td>
</tr>
<tr>
<td>204</td>
<td>S460</td>
<td>C-tail</td>
<td></td>
<td>Snr</td>
<td>0.06</td>
<td>4</td>
<td>0.87</td>
</tr>
<tr>
<td>205</td>
<td>V461</td>
<td>C-tail</td>
<td></td>
<td>Ivl</td>
<td>0.17</td>
<td>4</td>
<td>0.83</td>
</tr>
<tr>
<td>206</td>
<td>C463</td>
<td>C-tail</td>
<td></td>
<td>Rp</td>
<td>0.17</td>
<td>4</td>
<td>0.86</td>
</tr>
<tr>
<td>207</td>
<td>M466</td>
<td>C-tail</td>
<td></td>
<td>MV</td>
<td>0.69</td>
<td>4</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Sequence positions are indicated relative to the alignment (Align) as well as to Smad2, according to PDB 1KHX (29). Secondary structure elements are indicated according to Chen et al. (34). SH scores and corresponding size of sequential groups of low SH sites are listed. In addition, predictions for the other methods using default settings are shown (see text for details). The consensus patterns for the AR-Smads and BR-Smads are shown. All amino acid types are listed in order of decreasing frequency. Those of half or less than the frequency of the dominant type are in lower case. The known functions, with corresponding reference(s), are indicated. Putative functions corresponding to structural clustering shown in Figure 5A are indicated in brackets (see text for more detail).

Sequence retrieval and alignment

R-Smad protein sequences were collected using the NCBI query for sequence retrieval (www.ncbi.nih.gov). This resulted in 15 non-redundant sequences for AR-Smads and 17 for BR-Smads. All sequences were aligned using the PSI-Praline multiple sequence alignment online server (www.ibi.vu.nl/programs/pralinewww) (25,26). From the alignment obtained, the MH2 domain was selected for further analysis.
For the Ras superfamily, as used by Del Sol Mesa et al. (6),
sequences and alignments were directly obtained from Pfam-
B (27). Selection of sequences for Ras, Ral, Rab5 and Rab6
families was simply performed by matching sequence names
on Ras, Ral, Rab5 and Rab6, respectively yielding 69, 20,
4 and 6 sequences. The hypervariable termini of Rab5 and
Rab6 were not present in the Pfam alignment.

For the MIP family, we took all bacterial AQP and GLP
protein sequences, as defined in Figure 1 of Kalinina et al.
(7). For the other sequences mentioned in this figure, the
classification was less obvious. We therefore decided not to
take these into account. This scenario yielded 12 sequences
for the AQP subfamily and 48 for both GLP subfamilies
(one GLP identifier ‘YA17_HAEIN’ could not be resolved).
The sequences were aligned using PSI-Praline as mentioned
above for the Smads.

RESULTS

TGFβ-associated transcription factors (Smad family)
SH between AR- and BR-Smads was calculated using Equation
3 for all 211 positions along the Smad MH2 alignment,
as shown in Figure 1. It is clear that the vast majority is con-
served overall, i.e. 135 sites have SH = 1. On the other hand,
relatively few sites are completely non-harmonious, i.e. 32
have SH = 0. Out of these, only 13 are conserved in both
groups, as can been seen in Table 2. The other 19 sites
show a more variable composition but the subgroups still
have non-overlapping compositions. Further, a relatively
small fraction of sites show intermediate SH values, i.e. 44
are in between 0 and 1. In the range of 0–0.2, eight sites
are found, and between 0.2 and 0.8 only seven. This means
that the choice of a cut-off value for determining sites of
‘low’ SH is not very critical in this case.

Out of the 211 residues in the Smad MH2 sequence align-
ment, only 40 have a low SH (SH ≤ 0.2). In Table 2, these
sites are listed together with their known interactions. It is
clear that the vast majority of low-harmony sites also have
a known function. Table 3 provides a further summary of
the number of low-harmony sites associated with a particular
function. Out of the 40 low-harmony sites, 27 have a known
function (68%), while of the 32 non-harmonious sites
(SH = 0) 23 have a known function (72%). In total, there
are 29 known sites of functional specificity in the Smad
dataset. These include several with rather high compositional
variation in one or both groups (Table 2). Of the 171 remain-
ing sites (SH > 0.2), only two sites are known to be important
for the specificity of the Smad–receptor interactions (1%).

The AMAS method selects six sites that are different with
respect to one or more physicochemical properties (Table 2).
Consequently, these sites have non-overlapping amino acid
compositions between the groups. Three are conserved in
both groups, two are conserved in one group and one is not
conserved in either group. Five out of the six selected sites
are of known function, but clearly the majority of the 29
known functional sites are missed.

TreeDet/MB selects 21 sites that are completely conserved
in at least one group but show no intergroup overlap (Table 2).
Seven additional sites show the same characteristics but
are not selected. The 21 selected sites contain 16 known
functional sites. The other 7 sites contain 5 known functions.

SDP-pred selects 12 sites that are conserved in both groups
and show no overlap (Table 2). Of these, 9 have known func-
tions. One site of known function is conserved in both groups,
but is not selected by the method. SDP-pred reaches a max-
imum coverage of ~80%. All sites selected by SDP-pred are
also selected by TreeDet/MB (see above).

All sites selected by AMAS, TreeDet/MB or SDP-pred
are also selected by SH. SH selects 18 additional sites,
11 of which have a known function. None of the 18 sites
are conserved in both groups, which explains why the other
methods have difficulties finding them. Two further known
functional sites remain undetected by any of the included
methods.

The ROC plot shows that TreeDet/MB, SDP-pred and SH/E
all reach ~40% coverage with similar error (Figure 2). At
higher coverage, TreeDet and SH/E yield lower error than
SDP-pred. AMAS does not reach higher coverage and there-
fore it has not been applied to the other test-sets. It is clear
that SH outperforms its counterparts at all coverage/error
combinations. Notably, the first 14 sites selected by SH are
all validated functional sites.
Small GTPases (Ras superfamily)

In Figure 3, ROC-curves are shown for two sets of families from the Ras superfamily of small GTPases; Rab 5 versus Rab 6 (21,22), and Ras versus Ral (6,20).

For Rab5/6 specificity (Figure 3A), the SH predictions show a very high coverage, even at low error rates. Overall, TreeDet/MB, SDP-pred and SH/E achieve somewhat similar coverage and error. Nevertheless, SH/E reaches 35% coverage at only half the error of SDP-pred. At that coverage, TreeDet/MB shows a 2-fold error relative to SDP-pred. SH overall outperforms all other methods by a significant margin. The first 20 sites selected contain only two unknowns, while 70% coverage is attained within 10% error (Figure 3A).

For the Ras/Ral test-set, however, the prediction methods perform more similarly (Figure 3B). The first four sites (33%) selected by SDP-pred are validated. At higher coverage, SH/E performs slightly better than SDP-pred. Nevertheless, SDP-pred reaches 100% coverage at ~12% error, while at this error rate SH/E and SH are just above 80% coverage and TreeDet/MB attains 65% coverage.

Integral membrane transporters (MIP family)

Figure 4 shows the relative performance over the MIP family. Predictions with the SDP-pred server were obtained using our own alignment and are comparable with those reported by Kalinina et al. (7).

At 20% error, all methods only achieve medium coverage (~50%, Figure 4). In contrast, for the other test sets at 20% error all other methods achieve higher coverage (between 75% and 100%). SH achieves somewhat higher coverage at the lowest error rates (the first 3 selected sites are <5 Å from their nearest ligand). At higher error rates, SDP-pred generally outperforms the other methods. TreeDet/MB performs similar to SH at lower coverage and more similar to SDP-pred at higher coverage. For this dataset, ranking for SH was dominated by the harmony score and only minute differences were seen between SH and SH/E (data for SH/E not shown).

Spatial and functional clustering

In Figure 5, the SH data are projected onto representative crystal structures of the four protein families included in the benchmark.

For the Smad2 MH2 domain, we identified a limited number of spatial clusters of low-harmony sites as indicated in Figure 5A. Taking membership of these clusters as a guideline,
we assign putative functions to 10 out of the 13 unknowns, as indicated in Table 2. It is clear that most of these could not have been assigned from the sequence alone. Unknowns 392, 400, 407 and 410 can be assigned a putative function in FAST1 and/or Mixer binding. The two SARA-binding residues (366 and 368) in this FAST1/Mixer/SARA-binding cluster are furthest away from the unknowns. Importantly, 407 is known not to be involved in receptor interactions (28). Unknowns 334 and 337 can be assigned a putative function for Co-repressor (c-Ski/SnoN) binding. Position 337 is known not to be involved in SARA binding (29). Unknowns 269, 272, 273 and 443 can be assigned a putative function in SARA binding. In all, only three sites remain that cannot be assigned a putative function, out of a total of 40 low-harmony sites.

For the other test-sets, we have chosen layouts in Figure 5 analogous to those used in the corresponding papers (see Figure 5 caption). For Rab5/6 (Figure 5B) and Ras/Ral (Figure 5C), it can be seen that selected sites form localized groups in the structure. For the MIPs (Figure 5D), this trend is much less salient.

**DISCUSSION**

In this paper, we have introduced the SH as a means to pinpoint putative functionally different sites and compared the results with other methods that rely mostly on conservation.

**Performance varies for methods and datasets**

For the Smad test-set, all five methods reach >80% coverage at <10% error (Figure 2). Nonetheless, differences among the methods are substantial and SH maintains highest coverage virtually throughout. For the Rab5/6 test-set, all four methods perform at least moderately well (Figure 3A). The differences among the predictions are somewhat larger than for the Smad test-set, with coverage ranging from 40% to ~80% at 10% error. Also here, SH performs best.

As to the Ras/Ral test-set, all four methods perform well and reach >80% coverage at 10% error (Figure 3B). The differences among the predictions are small and there is no method that outperforms the others over a pronounced range in coverage or error.

For the MIP test-set, overall prediction quality attained by the four methods is dramatically lower than for the other test-sets, with coverages ranging from 20% to 40% at 10% error (Figure 4). Although the differences are relatively small, SH performs slightly worse at a coverage >35% than its counterparts for this dataset.

In summary, for two of the four test-sets all predictions are very accurate while for another set only SH achieves high accuracy. In contrast, for the fourth set none of the methods give accurate predictions. Furthermore, while predictions for two of the four test-sets show only small performance differences, the differences are larger for the remaining two test-sets and here SH achieves highest coverage throughout.
Different methods select different sites
AMAS focuses on conservation of physicochemical properties. This principle leads to very specific predictions but the existing method seems to be overly conservative. SDP-pred, TreeDet/MB and SH/E focus on conservation and yield good predictions in general. Differences among these methods are relatively small. This suggests that the signal arising from conservation dominates over possible differences arising from methodology or ranking schemes. However, there appears to be a margin for improvement that could indicate that other factors are more important than conservation for determining functional specificity.
SDP-pred uses the Bernoulli estimator to automatically determine an optimal cut-off and yields highly specific but conservative predictions. TreeDet uses an internal algorithm for unsupervised grouping of sequences that may not always lead to finding differences of interest.
SH focuses on non-overlapping composition between subfamilies and yields good predictions with very high coverage in several cases. This suggests that subfamily differences and sequence context are crucial determinants for functional specificity.
The examples studied here indicate that emphasis on conservation is not sufficient to specifically detect known functional sites. Shifting the focus completely to differences as we have implemented in the SH method, seems to give better predictions overall, at least on the datasets tested here. However, it remains to be seen whether other factors may be involved and what relative weight should be attached to conservation and compositional differences.
The difference between our SH and SH/E methods is the use of sequence context by SH for the ranking of selected sites. Generally, SH performs better than SH/E, which indicates that sequence context may be an important indicator to select regions of interest. It is interesting to note that several of the sequential ranges identified in the Smads (Table 2) are located inside helices. This may seem counter-intuitive since first and second neighbours are on opposite sides in a helix. Nevertheless, the average distance between the neighbouring Cα in a helix is only ~5 Å and the flexibility of the sidechain would allow them to approach close enough to participate in the same function. The alternating pattern of β-strands is accommodated by the inclusion of second neighbours, as already described. An interesting refinement of the method could integrate knowledge of the secondary structure, e.g. selecting only first neighbours for loops, only second neighbours for β-strands, and first, third or fourth neighbours for helices.

Performance varies for different datasets
For the protein families presented here, the functional specificities of many sites have been investigated. Validation based on point mutations (Smad, Ras/Ral) is the most specific, but ignores the possible cooperative role of additional sites. Validation based on the exchange of sequence segments (Rab5/6) does include the possible cooperative role of sites, but makes it difficult to assess the discriminatory effect of individual sites. Validation based on distance-to-ligand (MIP) provides no information about the individual role of sites. In all test-cases included, many sites remain that have not been investigated experimentally and it is likely that many additional functional sites exist. This would lead to an underestimation of the number of true positives and complicates the evaluation of different prediction results.
For the two test-sets based on point-mutants (Smad, Ras/Ral), we see a generally high performance of the methods. The abundance of direct and high quality experimental validation for the Smad test-set allows an accurate assessment of the quality of the predictions. For the other two test-sets (Rab5/6, MIP), performance is generally lower. The methods perform well in identifying specific sites, but experience more difficulty in delineating regions corresponding to swapped segments or residues that are close to bound ligands.
The degree of conservation differs significantly between sites, but the available experimental evidence does not necessarily cover this distribution uniformly. Notably, conserved sites are often likely to be functionally relevant and this has led to a likely bias in studying mutations of relatively conserved sites. Such a bias would lead to an overestimation of the importance of conservation.
For further development of this field of research, a well-crafted collection of high-quality experimental data for a variety of protein families would be of great value. Preferably, experimental validation would be based on a representative mix of sites with different degrees of conservation and include many specific point mutations as well as swapped segments to assess supporting roles of sites and strengthen confidence in true negatives.

Functional sites can be grouped by spatial clustering
Functionally related residues tend to form spatial clusters in a protein structure. SH selects sites of unknown function for the protein families considered, and most of these cluster with sites of known function. This provides a way of grouping selected sites and can be used to transfer functional annotation for residues assigned to the same cluster. Preliminary data seem to indicate that coarse-grained structures of medium or even low quality, e.g. by homology modelling or ab initio prediction, may also prove sufficient for this type of clustering.
For the Smad protein family many important questions about the specific interactions with other factors in the TGF-β and BMP-associated pathways are still open. The sites selected by SH are likely candidates for supporting these specific interactions. The putative functions assigned to these sites based on the spatial clustering may provide important guidance for future experiments.

CONCLUSION
We have shown that SH achieves predictions of high quality. While some other methods use sophisticated statistical analysis methods, this does not appear to lead to an increased quality of the predictions. The simplicity of SH makes the results easy to understand. SH achieves high coverage in general and selects additional sites of unknown function. The location of these sites in the crystal structures associated
with the benchmark sets used here indicates most of them as promising candidates for further investigation.

The current study provides, to the best of our knowledge, a first attempt of a systematic comparison of prediction methods for functional differences between protein subfamilies. From this analysis, we conclude that exploiting conservation alone is not sufficient, and that more emphasis on sequence differences and context could be the crucial factor for identification of sites of functional specificity.

The SH web-server is available at www.ibi.vu.nl/programs/seqharmwww.

ACKNOWLEDGEMENTS

K.A.F. and W.P. acknowledge financial support from the Netherlands Bioinformatics Centre, BioRange Bioinformatics research programme (SP 2.3.1 and SP 3.2.2, respectively). We thank the two anonymous referees for their thoughtful and thorough comments that have lead to significant improvements in the paper. We thank Willie Taylor for a critical reading of the manuscript. We also thank Mikhail Gelfand for sharing alignment data on the MIP family. Funding to pay the Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES