A Feature Selection Algorithm for Detecting Subtype Specific Functional Sites from Protein Sequences for SMAD Receptor Binding

Elena Marchiori*, Walter Pirovano, Jaap Heringa and Anton Feenstra*
* equally contributed
Centre for Integrative Bioinformatics VU (IBIVU)
Vrije Universiteit Amsterdam
The Netherlands
elena@few.vu.nl, pirovano@few.vu.nl, heringa@few.vu.nl, feenstra@few.vu.nl

Abstract

Multiple sequence alignments 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 (subtype specific sites). This paper proposes a new algorithm for selecting subtype specific sites from a set of aligned protein sequences. The algorithm combines a feature selection technique with neighbour position information for selecting and ranking a set of putative relevant sites. The algorithm is applied to a dataset of protein sequences from the MH2 domain of the SMAD family of transcription factors. Validation of the results on the basis of the known interaction and function of the sites shows that the algorithm successfully identifies the known (from literature) subtype specific sites and new putative ones.

1 Introduction

All known types of proteins are organized into families, and families are often again separated into subtypes on the basis of functional or genomic properties [1, 2]. As a consequence, methods have been introduced for the positional comparison of amino acid composition between groups of proteins from different families and/or subtypes (e.g., [3–6]).

More specifically, starting from a multiple sequence alignment (MSA) of the protein sub-families of interest, the aim is to identify sites that could be used to explain the (known) differences in function associated with these sub-families. We refer to this task as the “Subtype Specific Sites Selection” problem (the “4S problem” in short).

Conservation degree of aminoacid value at one site of a protein (sub)family has been considered as a key property for detecting subtype specific functional sites from protein sequences [4]. However, a complete characterization of subtype specific functional sites based on conservation is not yet available.

The following two conservation-based properties have been shown to be effective for detecting subtype specific sites:

(a) conservation within subfamilies but divergence among subfamilies. This property is used, e.g., in [7] for identifying ligand-binding functional sites;

(b) divergence among and within subfamilies. This property is used in [8] for identifying subtype specific functional sites.

The preferred measure to quantify conservation used in the majority of papers on detection of functional sites is Shannon (information) entropy. Typically, each site is evaluated independently by means of an entropy-based criterion, and the resulting distribution of values is used for identifying subtype specific functional sites [7].

In this paper we propose an alternative approach for the “4S problem” problem based on (multivariate) feature selection with domain knowledge.

First, we show that a popular feature relevance estimation algorithm, Relief, detects sites that satisfy property (a) (conservation within subfamilies but divergence among subfamilies).

Next, we introduce a straightforward extension of Relief to account also for property (b) (divergence among and within subfamilies) when measuring relevance of sites. The resulting algorithm assigns one score for each site estimating its subtype specific relevance. We use this scoring to select a subset of putative subtype specific sites.

Finally, we show how the resulting scoring can be improved by means of a novel heuristic. This heuristic is inspired by the observation of enhanced predictive performance reported in [8] when using the support of the signal...
by neighbouring selected subtype-specific sites for ranking sites with equal score. Here the score of each selected site (as computed by our Relief-based algorithm) is multiplied by a factor quantifying the “boosting” effect of its neighbour selected sites on the relevance of that site.

The resulting algorithm, called 4SA (for Subtype Specific Site Selection Algorithm), takes as input a set of protein sequences, applies a state-of-the-art MSA algorithm to align them, and outputs a subset of putative important sites and a scoring estimating their relevance as subtype specific sites.

We apply 4SA to the SMAD family of transcription factors. This family plays a crucial role in the transforming growth factor-β (TGF/β) signalling pathway, and is critical for determining 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 cell death. The TGF/β family of growth factors induce Type-I transmembrane receptors to phosphorylate and activate the receptor-regulated SMADs (R-SMADs) [9].

The R-SMADs can be subdivided into two major groups: the AR-SMADs which are mainly induced by TGF/β-type receptors (TB/βR-I), and the BR-SMADs which are mainly induced by the BMP-type receptors (BMPR-I). In addition, there are other types of receptors, like ALK1 and ALK2 that also activate BR-SMADs [10]. Subsequent association among SMADs is responsible for transport to the nucleus and the control of TGF/β target genes by association with and activation of transcription factor complexes.

The subclass specificity of the TGF/β and BMP pathways is well studied and therefore provides a wealth of experimental data for validation. On the other hand, there is still much to be learned about the specific interactions of the SMADs with other factors that determine the specificity of these pathways. This specificity is a crucial factor in separating the TGF/β and BMP associated pathways [9, 11]. It has been shown that most of the specific receptor-SMAD interactions, as well as the interactions with numerous additional proteins involved in this process, map to the so-called ‘Mad Homology 2’ (MH2) domain of the SMAD proteins, which constitutes one of the most conserved sequence regions [11].

We consider the specific dataset of protein sequences collected in [8], together with known (from the literature) subtype specific functional sites within the MH2 domain of the SMAD-family of transcription factors, which are used for validating our findings.

Using 4SA, we were able to identify all sites known to be associated with the TGF/β vs. BMP subclass specificity of SMADs. Moreover, we identified novel putative subtype specific sites.

2 Methods

The proposed algorithm considers a set of homologous protein sequences (represented by a sequence of aminoacids) which have been divided into two classes, e.g. according to receptor-binding specificity, like AR- and BR-SMADs. Our approach for detecting subtype specific sites consists of the following three main steps:

1) alignment. The sequences are aligned using a state-of-the-art MSA algorithm.

2) selection. The resulting aligned sequences are given as input to a feature selection algorithm, which selects and scores sites considered important.

3) scoring. The resulting scoring of the selected sites is refined using a knowledge-based heuristic.

Alignment

The alignment step is performed using the PSI-Praline MSA online server (see www.ibi.vu.nl/programs/pralinewww) [12, 13]. The output of alignment is a set of sequences of equal length built from the alphabet of aminoacids plus the ‘-’ (gap) symbol. Sequence sites become in this way well defined, where a site corresponds to one position in all sequences.

Selection

Site selection is tackled by means of an extension of Relief [14, 15], a popular technique for scoring features. Relief assigns a weight to features (here sites) based on how well the features separate samples from their nearest neighbours from the same and from the opposite class.

The algorithm constructs iteratively a weight vector, which is initially equal to zero. At each iteration Relief selects one sequence, adds to the weight the difference between that sequence and its nearest sequence from the opposite class (called nearest miss), and subtracts the difference between that sequence and its nearest neighbour from the same class (called nearest hit). The iterative process terminates when all sequences of the dataset have been considered. Subsampling can be used to improve efficiency in case of a large dataset. Pseudo-code of Relief for two classes of sequences is given below.

Relief

%input: X (two classes of aligned proteins)
%output: weights assigned to each site
%features are sites
%examples are sequences
nr_feat = total number of features;
weights = zero vector of size nr_feat;
for all exa in X do
  hit(exa) = nearest neighbour of exa

from same class;
miss(exa) = nearest neighbour of exa
from opposite class;
weights = weights-(exa-hit(exa))+(exa - miss(exa));
end;
return weights;

Here similarity of sequences is measured simply by the
Hamming distance. The difference between two sequences,
like (exa-hit(exa)), is a bit sequence, where each bit
represents the match (0) or mismatch (1) of aminoacids at
the corresponding site: for instance, ALM - BLM = 100.

A site will obtain best (maximum) weight if there is max-
imal local conservation within subfamilies (between each
pair of nearest neighbours of the same class) and maximal
local divergence among subfamilies (between each nearest
neighbour of opposite classes). Thus if a site satisfies prop-
erty (a) then its weight will be high.

We introduce the following procedure, called Diver-
sity which assigns weights to sites depending on their
contribution to the diversity between farthest neighbours of
opposite classes.

Diversity
%input: X (two classes of aligned proteins)
%output: weights assigned to each site
%features are sites
%examples are sequences
nr_feat = total number of features;
weights = zero vector of size nr_feat;
for all exa in X do
miss(exa) = farthest neighbour of exa
from opposite class;
weights = weights + (exa - miss(exa));
end;
return weights;

Diversity will assign maximum weight to a site
when all pairs of farthest neighbours of opposite classes
have different aminoacid values at that site. In particular,
if the site satisfies property (b) (divergence between sub-
families) then its weight will be high.

The resulting weights are added to those computed by Relief. We use a toy example illustrate the benefits of Relief+Diversity. Consider the following two classes of sequences of three aminoacids:

<table>
<thead>
<tr>
<th></th>
<th>s1</th>
<th>s2</th>
<th>s3</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>A</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>E</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>F</td>
<td>D</td>
</tr>
<tr>
<td>C2</td>
<td>A</td>
<td>A</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>B</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

| Relief: | -1 | 0  | 0  |
| Relief+Diversity: | 0  | 1  | 0  |

Application of Relief to C1,C2 yields (scaled) weight
-1 for s1, and 0 for both s2 and s3. Thus Relief consid-
ers s1 less relevant the s2 and s3, while the latter ones are
considered equally important (or, rather, equally irrelevant).

Application of Relief+Diversity yields equal
weight (= 0) for both site 1 and 3 , while weight of site
2 becomes equal to 1. Thus now both s1 and s3 are judged
as equally (un)important, while site 2 is considered more
important, because even if it is not conserved within each
class, its divergence between opposite classes can be used
for discriminating C1 and C2.

So Relief+Diversity considers one site relevant
even when it is not conserved within subfamilies, provided
it is divergent between subfamilies. Then application of Relief+Diversity will yield high weights for sites satisfying property (a) or (b).

Sites with weight greater than or equal to 0.5 are selected
as putative subtype specific sites. This happens when, for
at least half of the sequences in the dataset, there is a match
with nearest neighbour of the same class and a mismatch
with nearest neighbour of opposite class.

Scoring

Suppose sites selected by an algorithm are contained in
a range of consecutive sequence positions separated from
each other by at most one position. The size of this range
can be used as value for boosting relevance (weight) of the
selected sites with positions in that range, by multiplying
their respective weights by that value. We call this heuristic
Context.

The following toy example illustrates application of this
heuristic. Suppose Relief+Diversity selected the fol-
lowing sites:

<table>
<thead>
<tr>
<th>site</th>
<th>index</th>
<th>position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relief+Diversity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relief+Diversity+Context</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R+D</td>
<td></td>
<td></td>
</tr>
<tr>
<td>w1</td>
<td></td>
<td>w1+w2</td>
</tr>
<tr>
<td>w2</td>
<td></td>
<td>w2+w3</td>
</tr>
<tr>
<td>w3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>w4</td>
<td></td>
<td>w4+w5+w6</td>
</tr>
<tr>
<td>w5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>w6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>w7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>w8</td>
<td></td>
<td>w8+w9+2</td>
</tr>
<tr>
<td>w9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The last two columns indicate site weights computed
with Relief+Diversity and Relief+Diversity+ Context, respectively. Consecutive sites in the range 1:5
and in range 77:78 are separated by at most one position.
So their Relief+Diversity weight will be multiplied
by the size of their respective ranges, that is, 5 and 2.

The final algorithm Relief+Diversity+Context
is called 4SA. It applies sequentially Relief, Diver-
sity and Context. Its output is a subset of selected sites and a scoring of all sites, where weights of non-selected sites (considered irrelevant) are set to a value lower than the one of any of the selected sites.

3 Data

R-SMAD protein sequences were collected using the NCBI query for sequence retrieval (www.ncbi.nih.gov). The resulting 32 non-redundant R-SMAD sequences were aligned using PSI-Pralline MSA (www.ibivu.cs.vu.nl/programs/pralinewww) [12, 13]. The MH2 domain was selected for further analysis; this is non-redundant for 8 AR-SMADs and for 12 BR-SMADs. The alignment was divided into two subgroups according to receptor-binding specificity: AR-SMADs (SMADs 2 and 3, binding TB/β-R-1) and BR-SMADs (SMADs 1, 5 and 8, binding BMPR-I/ALK1/2).

3.1 Evaluation

The performance of our method is tested against the known subtype specific sites for the SMAD MH2 domain, using a dataset previously assembled in [8]. The set consists of 29 sites that were specifically determined to be involved in changing the specificity of AR- and BR-SMADs for either receptor type.

4 Results

Receiver-operator characteristic (ROC) curve provides a tool for assessing the performance of an algorithm [16, 17]. Here known subtype specific sites are considered (true) positive, and the unknown sites are considered (true) negatives. We use the scoring (weight) values as threshold for generating the ROC curve. For each weight value v the set of sites with weight higher than or equal to v is considered: the true positive percentage is reported on the y-axis (sensitivity), and the false positive percentage (1-specificity) on the x-axis.

Observe that the best possible ROC curve will be obtained by a method that selects the set of known sites. It can be argued that this is a conservative assessment of the quality of a method for detecting subtype specific sites. Unfortunately, from the available experimental evidence in literature, it is impossible to identify false positive subtype specific site detections, and, likewise, no direct evidence is present to discriminate true from false negatives. Therefore here the ROC curve describes the goodness of a method in giving higher ranking to the known subtype specific sites.

We applied the following four Relief-based algorithms to the R-SMAD dataset: Relief, Relief+Diversity, Relief+Context, and the complete algorithm 4SA. Figure 1 contains the corresponding ROC curves, indicating that improved performance is achieved when using all the components of 4SA.

Figure 1. ROC curves of Relief-based algorithms for subtype specific sites detection.

5 Discussion

Table 1 contains details of 47 sites selected by 4SA, including a summary of all literature data on subtype specificity and results of other state-of-the-art algorithms applied to this dataset.

From literature, 29 sites are known to be important for receptor-type specificity. In addition to being directly involved in TB/β-R-1 or BMPR-I binding, this includes sites involved in other pathway specific factors, like transcription factors, cytosolic retention factors (FAST1, SARA or Mixer) and co-repressors (C-Ski/SnoN). The 47 sites selected by 4SA contain all 29 known sites.

For the remaining 18 sites of unknown function that were selected by 4SA, we have assigned putative functions based on proximity in the SMAD2 crystal structure 1KHX [18]. Figure 2 shows the crystal structure with the 4SA sites highlighted. Putative functions were taken from the closest site of known function. However, if a putative function for a closer site of unknown function disagrees with the function of the known site, no putative function was assigned. This was the case for only 2 sites. For 16 sites out of 18 of unknown function putative functions could be assigned unequivocally.
Figure 2. 4SA site selection for AR- vs. BR-SMADs colour-coded onto the crystal structure of the MH2 domain of SMAD2 (1KHX [18]), cf. Table 1. 4SA selected sites are in red, shown with sidechains and labelled with residue name and number.

6 Comparison with Other Methods

Current approaches to the 4S problem are diverse, such as those based on entropy (2-Entropies [7]), (Sequence Harmony [8]), relative entropy and Mutual Information and Bernoulli estimators (e.g. SDP-pred [19]), tree-determinant sites (e.g. TreeDet [4]), and hierarchical conservation of physicochemical properties of amino acids (e.g. AMAS [20]).

We compare by means of ROC plots 4SA and four state-of-the-art algorithms: Sequence Harmony [8], SDP-pred [19, 21], TreeDet [4], and AMAS [20].

For these algorithms, the following online servers have been used: SDP-pred math.genebee.msu.ru/~psn, TreeDet somosierra.cnb.uam.es/Servers/treedetv2, AMAS barton.ebi.ac.uk/servers/amas_server.html, SH www.ibi.vu.nl/programs/segharmwww.

We consider the results reported in [8], which use the default parameter settings of these algorithms were used (i.e., the most significant Z-score cutoff using Bernoulli estimators for SDP-pred, a cutoff of .6 and 10% high-scoring residues for TreeDet, a conservation threshold of 7 for AMAS, and maximal sequence harmony equal to .2).

The best ROC curve is obtained by 4SA, indicating improved capability of detecting subtype specific sites for the SMAD receptor binding.

7 Conclusion

In this paper we showed that feature selection and scoring performed by a Relief-based algorithm with residue context information successfully identifies subtype specific sites for the binding of SMAD to the receptor and other TGFβ pathway components.

All SMAD MH2 selected sites of known function form functionally related clusters in the MH2 domain structure. In addition, 4SA identified 18 sites for which to the best of our knowledge the function has not been experimentally verified. Using spatial proximity with sites of known function, we suggest putative functions for 16 of these sites of unknown function. Based on our 4SA analysis of the SMAD MH2 sequences, the differences in amino acid consensus patterns observed for selected sites and their spatial arrangement, we suggest these sites to be interesting candidates for further (experimental) study.

References


Table 1. Summary of all known functional sites and sites detected by 4S, with corresponding 4S ranking, and results for four other prediction methods, in the MH2 domain of Smad. Sequence positions are indicated relative to the alignment and Smad2, according to PDB 1KHX [18]. Consensus patterns for the AR-Smads and BR-Smads are shown, with all amino acid types listed in order of decreasing frequency, and those of half or less than the frequency of the dominant type in lower case. Known interactions, putative functions, and corresponding literature references are listed.

<table>
<thead>
<tr>
<th>Position</th>
<th>Consensus</th>
<th>Algorithms</th>
<th>Interaction (putative)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Align</td>
<td>SMAD2</td>
<td>AR BR 4SA</td>
<td>SH SDP-pred AMAS Tree-det</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>(L263) La Vfm</td>
<td>16 0 - - + -</td>
<td>-</td>
<td>SARA</td>
</tr>
<tr>
<td>3</td>
<td>(Q264) Qa Qrh</td>
<td>46 - - - -</td>
<td>-</td>
<td>SARA</td>
</tr>
<tr>
<td>6</td>
<td>T267 Tm Acen</td>
<td>19 0 - - -</td>
<td>-</td>
<td>SARA</td>
</tr>
<tr>
<td>8</td>
<td>S269 CSh Eq</td>
<td>20 0 - - -</td>
<td>-</td>
<td>(SARA)</td>
</tr>
<tr>
<td>11</td>
<td>A272 A Kqls</td>
<td>25 0 - - -</td>
<td>-</td>
<td>(c-Ski/SnoN)</td>
</tr>
<tr>
<td>12</td>
<td>F273 F Hy</td>
<td>23 0 - - -</td>
<td>-</td>
<td>(c-Ski/SnoN)</td>
</tr>
<tr>
<td>17</td>
<td>A278 SA Va</td>
<td>38 - - - +</td>
<td>-</td>
<td>(—)</td>
</tr>
<tr>
<td>23</td>
<td>Q284 Qt N</td>
<td>34 0 - - -</td>
<td>-</td>
<td>TB/3R-1</td>
</tr>
<tr>
<td>28</td>
<td>T289 T At</td>
<td>41 - - - -</td>
<td>-</td>
<td>(—)</td>
</tr>
<tr>
<td>33</td>
<td>Q294 Q Sq</td>
<td>11 .16 - - -</td>
<td>-</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>34</td>
<td>P295 P Trl</td>
<td>5 0 - - .85</td>
<td>-</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>36</td>
<td>L297 LMi Vi</td>
<td>13 .11 - - -</td>
<td>-</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>37</td>
<td>T298 T Li</td>
<td>6 0 - - .88</td>
<td>-</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>47</td>
<td>S308 Sa N</td>
<td>15 0 - - -</td>
<td>-</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>48</td>
<td>- - Nsd</td>
<td>21 0 - - -</td>
<td>-</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>49</td>
<td>E309 E KrS</td>
<td>17 0 - - -</td>
<td>-</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>63</td>
<td>A323 Ae S</td>
<td>7 0 - - .84</td>
<td>-</td>
<td>ALK1/2</td>
</tr>
<tr>
<td>64</td>
<td>T324 ATr T</td>
<td>18 - - - -</td>
<td>-</td>
<td>(ALK1/2)</td>
</tr>
<tr>
<td>65</td>
<td>V325 V I</td>
<td>1 0 2.17 -.87</td>
<td>-</td>
<td>ALK1/2</td>
</tr>
<tr>
<td>67</td>
<td>M327 Lmq N</td>
<td>8 0 - + .83</td>
<td>-</td>
<td>ALK1/2</td>
</tr>
<tr>
<td>74</td>
<td>R334 Rk K</td>
<td>42 .18 - - -</td>
<td>-</td>
<td>(c-Ski/SnoN)</td>
</tr>
<tr>
<td>77</td>
<td>R337 R H</td>
<td>26 0 2.25 -.87</td>
<td>not SARA (c-Ski/SnoN)</td>
<td>[18]</td>
</tr>
<tr>
<td>81</td>
<td>I341 I V</td>
<td>27 0 2.24 -.87</td>
<td>-</td>
<td>sara/Mixer</td>
</tr>
<tr>
<td>86</td>
<td>F346 F Y</td>
<td>28 0 2.14 -.87</td>
<td>-</td>
<td>sara/Mixer</td>
</tr>
<tr>
<td>94</td>
<td>A354 As S</td>
<td>43 .18 - - -</td>
<td>-</td>
<td>(SARA/Mixer)</td>
</tr>
<tr>
<td>100</td>
<td>P360 P R</td>
<td>29 0 2.21 + .87</td>
<td>-</td>
<td>FAST1</td>
</tr>
<tr>
<td>104</td>
<td>Q364 Q Yf</td>
<td>9 0 - - -</td>
<td>-</td>
<td>Mixer/FAST1</td>
</tr>
<tr>
<td>105</td>
<td>R365 R Hq</td>
<td>2 0 - - -</td>
<td>-</td>
<td>Mixer/FAST1</td>
</tr>
<tr>
<td>106</td>
<td>Y366 Y H</td>
<td>3 0 2.02 -.86</td>
<td>-</td>
<td>sara/Mixer/FAST1</td>
</tr>
<tr>
<td>108</td>
<td>W368 W F</td>
<td>4 0 2.22 -.87</td>
<td>-</td>
<td>sara/Mixer/FAST1</td>
</tr>
<tr>
<td>111</td>
<td>A371 Ta -</td>
<td>44 - - -</td>
<td>-</td>
<td>(FAST1)</td>
</tr>
<tr>
<td>118</td>
<td>P378 P Sp</td>
<td>39 .16 - - -</td>
<td>-</td>
<td>(SARA/Mixer)</td>
</tr>
<tr>
<td>121</td>
<td>N381 N S</td>
<td>30 0 - .87</td>
<td>-</td>
<td>sara/Mixer</td>
</tr>
<tr>
<td>136</td>
<td>A392 A Qeh</td>
<td>37 0 - - .85</td>
<td>-</td>
<td>(Mixer/FAST1)</td>
</tr>
<tr>
<td>143</td>
<td>N399 Nh Ns</td>
<td>45 - - - -</td>
<td>-</td>
<td>(SARA/Mixer/FAST1)</td>
</tr>
<tr>
<td>144</td>
<td>Q400 Q H</td>
<td>22 0 2.03 + .87</td>
<td>-</td>
<td>(Mixer/FAST1)</td>
</tr>
<tr>
<td>151</td>
<td>Q407 Qr E</td>
<td>35 0 - .83</td>
<td>-</td>
<td>not receptor binding (FAST1)</td>
</tr>
<tr>
<td>154</td>
<td>R410 R K</td>
<td>31 0 2.28 -.87</td>
<td>-</td>
<td>(FAST1)</td>
</tr>
<tr>
<td>171</td>
<td>R427 R H</td>
<td>32 0 2.01 -.87</td>
<td>-</td>
<td>TB/3R-1/BMPR-1/ALK1/2</td>
</tr>
<tr>
<td>174</td>
<td>T430 T D</td>
<td>33 0 2.08 -.87</td>
<td>-</td>
<td>TB/3R-1/BMPR-1/ALK1/2</td>
</tr>
<tr>
<td>184</td>
<td>L440 L Iv</td>
<td>36 0 - - .84</td>
<td>-</td>
<td>(c-Ski/SnoN)</td>
</tr>
<tr>
<td>187</td>
<td>N443 N Hn</td>
<td>40 .16 - - -</td>
<td>-</td>
<td>(c-Ski/SnoN)</td>
</tr>
<tr>
<td>204</td>
<td>S460 Snr Hlr</td>
<td>14 .06 - - -</td>
<td>-</td>
<td>TB/3R-1/BMPR-1</td>
</tr>
<tr>
<td>205</td>
<td>V461 Ivl N</td>
<td>12 0 - + .83</td>
<td>-</td>
<td>TB/3R-1/BMPR-1</td>
</tr>
<tr>
<td>206</td>
<td>R462 Rp P</td>
<td>24 .17 - - -</td>
<td>-</td>
<td>TB/3R-1/BMPR-1</td>
</tr>
<tr>
<td>207</td>
<td>C463 C I</td>
<td>10 0 2.30 + .86</td>
<td>-</td>
<td>TB/3R-1/BMPR-1</td>
</tr>
<tr>
<td>210</td>
<td>M466 VM V</td>
<td>47 - - - -</td>
<td>-</td>
<td>TB/3R-1/BMPR-1</td>
</tr>
</tbody>
</table>

Total selected: 47 40 12 6 21
selected functional: 29 27 8 4 14
selected putative: 16 12 4 1 –
selected unknown: 2 0 4 1 7

29 known functional sites in total