SAAPpred: A New Server For Predicting The Pathogenicity of Mutations

Nouf S. Al-Numair[a,c], David S. Gregory[b], Andrew C. R. Martin[a,*]

[a] *Institute of Structural and Molecular Biology,Division of Biosciences, University College London, Gower Street, London WC1E 6BT.*
[b] *Division of Biosciences and Department of Computer Science, University College London, Gower Street, London WC1E 6BT.*
[c] *Present address: Department of Genetics,Research Centre King Faisal Specialist Hospital and Research Centre MBC-03, PO Box 3354, Riyadh 11211, Saudi Arabia*

Abstract

We present the SAAPpred server, a prediction server for assessing whether single amino acid mutations are likely to be pathogenic. SAAPpred is built on SAAPdap, our data analysis pipeline which analyses a set of structural features. While SAAPdap offers information about individual features that may be damaging (for example a small-to-large mutation causing a clash), SAAPpred amalgamates the results of all the analyses and uses a random forest to predict whether a mutation is pathogenic. The underlying resources used for the analysis and prediction have been updated and are now automatically updated on a regular basis.

The server is available at `www.bioinf.org.uk/saap/dap/`.

Keywords: Pathogenic deviations; Interpreting mutation data

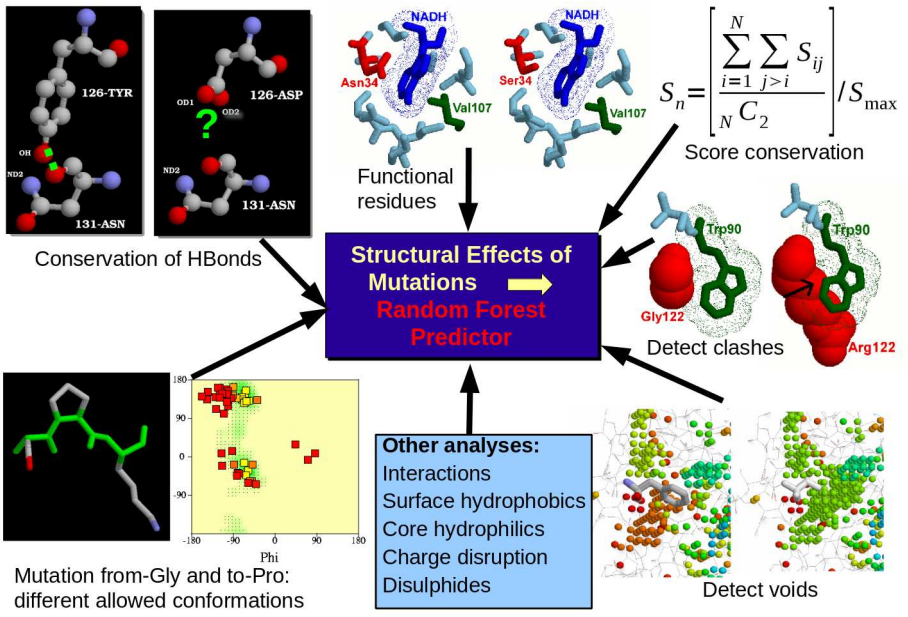
1. Research highlights

- We present a new web server to access our SAAPpred pathogenicity prediction system.
- The analysis and prediction method relies on having a protein structure available for the protein in question.

*Corresponding author

- The resource has been updated with newly available PDB structures and sequence families for calculation of conservation information. Updating is now automated.
- Performance appears to exceed that of competing methods for pathogenicity prediction.

2. Graphical Abstract



3. Introduction

High-throughput next-generation sequencing platforms[1] are increasingly used to screen patients with genetic disease for pathogenic mutations. This has led to a huge demand for methods that can analyze and predict the effects of mutations, but prediction remains challenging.

Most mutations are 'loss of function' although some are 'gain of function' (generally through loss of regulation). A small number are actually 'change of function', for example of specificity, and it has been estimated that 5% of cancer mutations fall into this category[2]. At least 20 research groups

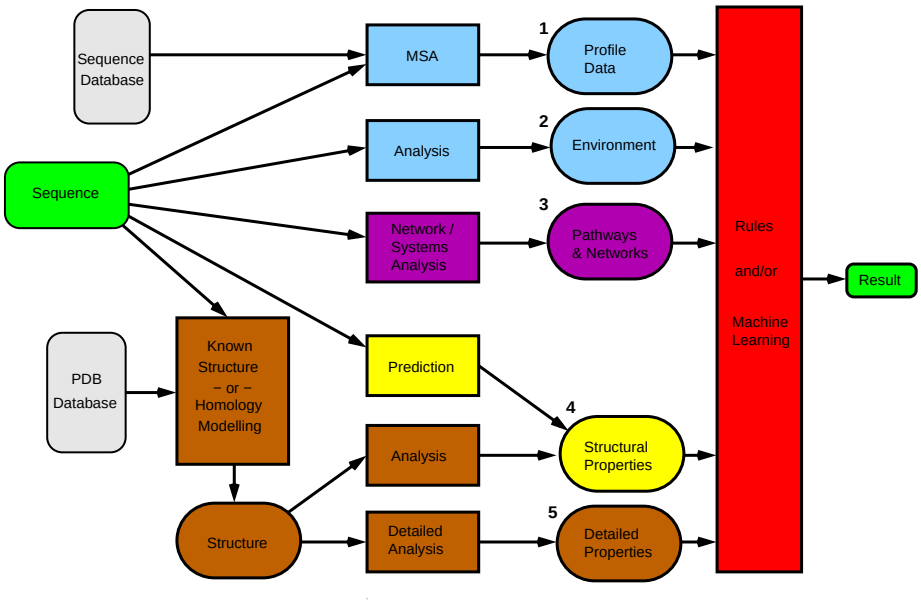


Figure 1: Classification of approaches used to predict the effect of mutations.

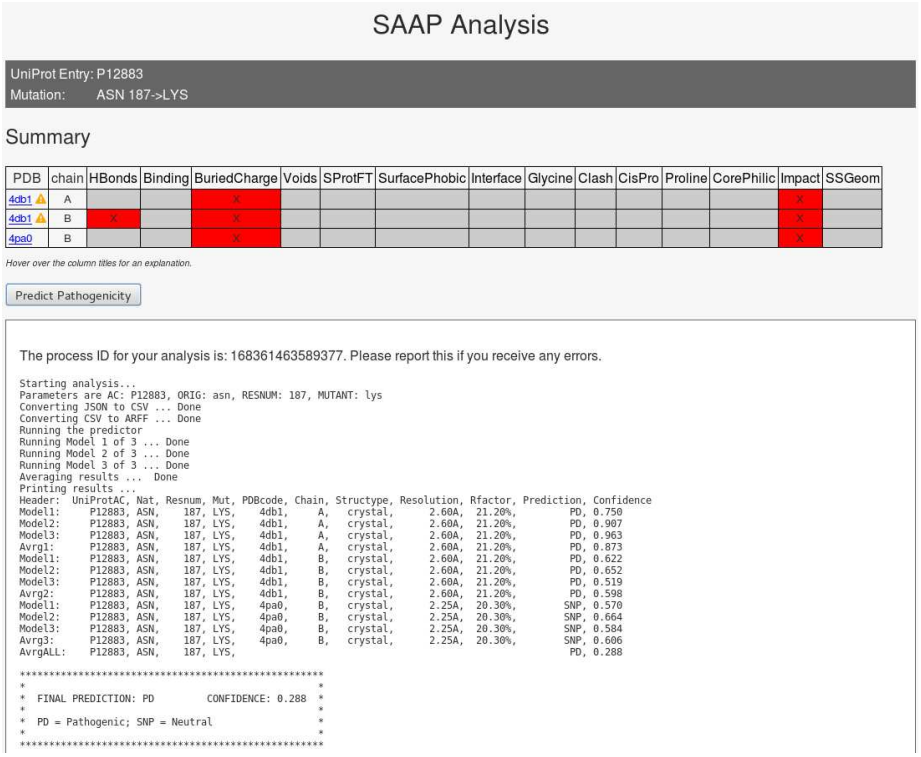
have developed prediction methods which adopt a variety of approaches and we have attempted to provide a reasonably comprehensive list at `www.bioinf.org.uk/saap/methods/`.

Methods can be classified based on the types of data they use (sequence, structure and/or systems biology / network information) and on the methods that they use to interpret the information (rules, machine-learning, etc.). Figure 1 illustrates the types of approaches. Many methods (boxes 1 and 2 in Figure 1) are purely sequence based, typically including evolutionary conservation information. For example, SIFT[3], Align-GVGD[4], MutationAssessor[2], PANTHER[5], MAPP[6] and FATHMM[7]. Other methods add structural information (box 4), either directly from a known structure or predicted structural properties such as secondary structure or solvent accessibility (for example, PolyPhen[8], PolyPhen-2[9], LS-SNP/PDB[10], SNPeffect[11], BONGO[12], SNAP[13], PMUT[14] and CanPredict[15]). When a structure is not available, some methods also exploit comparative modelling (e.g. LS-SNP[16]) or *ab initio*

structural models[17]. Methods such as SuSPect[18] also incorporate systems biology-based features (box 3). Some methods such as SNP@Domain[19] and SNPs3D[20] rely more heavily on structure (box 5) including features such as hydrogen-bonding, clashes and voids. Our own SAAPdb[21], a pre-calculated analysis database, more recently replaced by SAAPdap[22], an on-the-fly data analysis pipeline that caches its results, together with SAAPpred[22], our prediction system, fall into this category including analysis of clashes and voids, conservation of hydrogen bonds and a Ramachandran pseudo-energy for mutations to proline or from glycine.

Methods used to interpret the information include empirical rules (e.g. PolyPhen[8]), direct methods (exploiting a score based on some type of theoretical model of what happens when a mutation occurs — e.g. SIFT[3] and PANTHER[5]) and, most commonly, machine-learning methods. These include artificial neural networks, support vector machines and random forests, and can combine different properties of the native and mutant residue such as size and polarity, together with other information such as structural environment (e.g. accessibility and hydrogen bonding) as well as evolutionary conservation. Examples include PMUT[14], SNAP[13], PhD-SNP[23], SNPs&GO[24], Parepro[25], CanPredict[15], nsSNPAnalyzer[26], MutPred[27], Hansa[28] and MutationTaster[29] as well as our own SAAPpred[22]. More recent additions to the list of approaches include consensus predictors such as Condel[30], PredictSNP[31] and DUET[32].

Almost all methods are general predictors which can be applied to any mutation although some, like CanPredict, are specialized for mutations in cancer. In general, while the overall performance of a method may be good, the performance on a particular protein or disease is rarely evaluated. However, one method, KvSNP[33] has been designed specifically for analysis of mutations in voltage-gated potassium channels.



SAAP Analysis

UniProt Entry: P12883
Mutation: ASN 187->LYS

Summary

PDB	chain	HBonds	Binding	BuriedCharge	Voids	SProtFT	SurfacePhobic	Interface	Glycine	Clash	CisPro	Proline	CorePhilic	Impact	SSGeom
4db1	A			X										X	
4db1	B	X		X										X	
4pa0	B			X										X	

Hover over the column titles for an explanation.

Predict Pathogenicity

The process ID for your analysis is: 168361463589377. Please report this if you receive any errors.

```
Starting analysis...
Parameters are AC: P12883, ORIG: asn, RESNUM: 187, MUTANT: lys
Converting JSON to CSV ... Done
Converting CSV to ARFF ... Done
Running the predictor
Running Model 1 of 3 ... Done
Running Model 2 of 3 ... Done
Running Model 3 of 3 ... Done
Averaging results ...  Done
Printing results ...
Header:  UniProtAC, Nat, Resnum, Mut, PDBcode, Chain, Structype, Resolution, Rfactor, Prediction, Confidence
Model1:     P12883, ASN,    187, LYS,    4db1,     A,   crystal,      2.60A,  21.20%,         PD, 0.750
Model2:     P12883, ASN,    187, LYS,    4db1,     A,   crystal,      2.60A,  21.20%,         PD, 0.907
Model3:     P12883, ASN,    187, LYS,    4db1,     A,   crystal,      2.60A,  21.20%,         PD, 0.963
Avrg1:      P12883, ASN,    187, LYS,    4db1,     A,   crystal,      2.60A,  21.20%,         PD, 0.873
Model1:     P12883, ASN,    187, LYS,    4db1,     B,   crystal,      2.60A,  21.20%,         PD, 0.622
Model2:     P12883, ASN,    187, LYS,    4db1,     B,   crystal,      2.60A,  21.20%,         PD, 0.652
Model3:     P12883, ASN,    187, LYS,    4db1,     B,   crystal,      2.60A,  21.20%,         PD, 0.519
Avrg2:      P12883, ASN,    187, LYS,    4db1,     B,   crystal,      2.60A,  21.20%,         PD, 0.598
Model1:     P12883, ASN,    187, LYS,    4pa0,     B,   crystal,      2.25A,  20.30%,        SNP, 0.570
Model2:     P12883, ASN,    187, LYS,    4pa0,     B,   crystal,      2.25A,  20.30%,        SNP, 0.664
Model3:     P12883, ASN,    187, LYS,    4pa0,     B,   crystal,      2.25A,  20.30%,        SNP, 0.584
Avrg3:      P12883, ASN,    187, LYS,    4pa0,     B,   crystal,      2.25A,  20.30%,        SNP, 0.606
AvrgALL:    P12883, ASN,    187, LYS,                                                         PD, 0.288

**************************************************
*                                                *
*  FINAL PREDICTION: PD         CONFIDENCE: 0.288  *
*                                                *
*  PD = Pathogenic; SNP = Neutral                *
*                                                *
**************************************************
```

Figure 2: The SAAPdb/SAAPpred web site showing prediction on an Asn187Lys mutation in the protein Myosin-7 which is known to be pathogenic.

4. Results

Our own approach was initially to try to understand the local structural effects caused by mutations, comparing these effects in single nucleotide polymorphisms (SNPs, that is non-pathogenic mutations) and pathogenic deviations (PDs). In many cases, these structural effects (together with a measure of conservation) could be used to suggest that a mutation would be damaging. We then went on to use these analysis results to train a random forest machine learning method. As described previously[22], the performance of the method, on the set of mutations for which it can be used — those that occur in proteins for which a structure is available (MCC=0.692), exceeds other commonly used methods such as SIFT (MCC=0.528), PolyPhen-2 (MCC=0.572), MutationAssessor (MCC=0.453) and FATHMM (MCC=0.671) evaluated on the same dataset. It should also be noted that our results are fully cross-validated; not only do we not allow the same mutation to appear in the training and testing sets, but also we do not allow the same protein to appear in both datasets. In Al-Numair & Martin[22] we also showed a non-cross-validated result of MCC=0.894. In fact, even this set did not allow the same mutation in the training and testing sets although it did allow the same structure to appear in both datasets.

Because SAAPpred relies on the SAAPdap analysis of the effects of mutations which requires protein structures from the Protein Databank (PDB) we now automatically update the available PDB data on a daily basis. In addition we have updated FOSTA, our database of functionally equivalent orthologues[34] which is used to calculate conservation information. FOSTA contains families of protein sequences each of which is rooted around a human sequence. The previous release of FOSTA contained 535697 sequences in 20245 families while the new version contains 551384 sequences in 20197 families. The decrease in the number of families corresponds to fewer human sequences being present in UniProKB/SwissProt. It is not clear why these sequences have been removed by the curators. Neither it is clear which sequences have

been removed since FOSTA uses UniProKB/SwissProt IDs as keys and IDs can, and do, change.

In addition, we have developed a web interface to allow the SAAPpred predictor to be run, as shown in Figure 2. Users enter the UniProtKB/SwissProt accession code, the residue number and the native and mutant amino acids in order to run the SAAPdap pipeline. Once this completes, the user is presented with a summary results table, below which is a 'Predict Pathogenicity' button. Clicking this expands a panel where progress on the pathogenicity prediction is shown. The final results of a prediction are shown in the figure. Once SAAPdap has completed, SAAPpred takes up to two minutes to complete.

5. Materials and Methods

PDB files are mirrored using 'FTPMirror' available as part of the 'bioscripts' package (`github.com/AndrewCRMartin/bioscripts`). FTPMirror automatically handles decompression of remote compressed files and can handle very large remote directories containing > 65536 files (such as the PDB) which fail with the Sunsite Mirror script and the Perl LWP package. FOSTA updates are performed on a heterogeneous network of Linux machines using the Sun Grid Engine.

The web server is implemented in Perl using the Weka machine learning environment as described in Al-Numair & Martin[22]. The web site follows XHTML1.1 standards with CSS formatting and uses AJAX to keep the user updated with progress until the results are complete. Results are calculated over three prediction models and up to three PDB chains. PD (damaging) and SNP (phenotypically silent) predictions are averaged separately and the difference in the averages is presented as the final prediction.

6. Conclusion

This update and interface to the SAAPpred server adds to the range of tools available for evaluation of the pathogenicity of a single amino acid mutation. In

our hands, the method out-performs other well-known methods including SIFT and PolyPhen-2 as well as more recent methods such as FATHMM. The server is available at `www.bioinf.org.uk/saap/dap/`.

7. Acknowledgements

NSAN thanks the King Faisal Specialist Hospital and Research Centre and the Royal Embassy of Saudi Arabia Cultural Bureau (reference S12063) for funding.

References

[1] D. R. Bentley, Whole-genome re-sequencing, Curr Opin Genet Dev 16 (2006) 545–552.

[2] B. Reva, Y. Antipin, C. Sander, Predicting the functional impact of protein mutations: Application to cancer genomics, Nucleic Acids Res 39 (2011) e118–e118.

[3] P. C. Ng, S. Henikoff, SIFT: predicting amino acid changes that affect protein function, Nucleic Acids Res 31 (2003) 3812–3814.

[4] E. Mathe, M. Olivier, S. Kato, C. Ishioka, P. Hainaut, S. V. Tavtigian, Computational approaches for predicting the biological effect of p53 missense mutations: A comparison of three sequence analysis based methods, Nucleic Acids Res. 34 (2006) 1317–25.

[5] P. D. Thomas, M. J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, A. Narechania, PANTHER: A library of protein families and subfamilies indexed by function., Genome Research 13 (2003) 2129–2141.

[6] J. Binkley, K. Karra, A. Kirby, M. Hosobuchi, E. A. Stone, A. Sidow, ProPhylER: a curated online resource for protein function and structure based on evolutionary constraint analyses, Genome Res 20 (2010) 142–154.

[7] H. A. Shihab, J. Gough, D. N. Cooper, P. D. Stenson, G. L. A. Barker, K. J. Edwards, I. N. M. Day, T. R. Gaunt, Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models, Hum Mutat 34 (2013) 57–65.

[8] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, S. R. Sunyaev, A method and server for predicting damaging missense mutations., Nature Methods 7 (2010) 248–249.

[9] I. A. Adzhubei, D. M. Jordan, S. R. Sunyaev, Predicting functional effect of human missense mutations using PolyPhen-2, Curr Protoc Hum Genet 76 (2013) 7.20.

[10] M. Ryan, M. Diekhans, S. Lien, Y. Liu, R. Karchin, LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures, Bioinformatics 25 (2009) 1431–2.

[11] J. Reumers, S. Maurer-Stroh, J. Schymkowitz, F. Rousseau, SNPeffect v2.0: A new step in investigating the molecular phenotypic effects of human non-synonymous SNPs, Bioinformatics 22 (2006) 2183–2185.

[12] T. M. K. Cheng, Y.-E. Lu, M. Vendruscolo, P. Lio, T. L. Blundell, Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms, PLoS Comp. Biology 4 (2008) e1000135.

[13] Y. Bromberg, G. Yachdav, B. Rost, SNAP predicts effect of mutations on protein function, Bioinformatics 24 (2008) 2397–2398.

[14] C. Ferrer-Costa, J. L. Gelp, L. Zamakola, I. Parraga, X. de la Cruz, M. Orozco, PMUT: a web-based tool for the annotation of pathological mutations on proteins, Bioinformatics 21 (2005) 3176–8.

[15] J. S. Kaminker, Y. Zhang, C. Watanabe, Z. Zhang, CanPredict: a computational tool for predicting cancer-associated missense mutations, Nucleic Acids Res. 35 (2007) W595–W598.

[16] R. Karchin, M. Diekhans, L. Kelly, D. J. Thomas, U. Pieper, N. Eswar, D. Haussler, A. Sali, LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources, Bioinformatics 21 (2005) 2814–2820.

[17] C. T. Saunders, D. Baker, Evaluation of structural and evolutionary contributions to deleterious mutation prediction, J Mol Biol 322 (2002) 891–901.

[18] C. M. Yates, I. Filippis, L. A. Kelley, M. J. Sternberg, SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features, J Mol Biol 426 (2014) 2692–2701.

[19] D. Chasman, R. M. Adams, Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: Structure-based assessment of amino acid variation, J Mol Biol 307 (2001) 683–706.

[20] P. Yue, E. Melamud, J. Moult, SNPs3D: candidate gene and SNP selection for association studies, BMC Bioinformatics 7 (2006) 166–166.

[21] J. M. Hurst, L. E. M. McMillan, C. T. Porter, J. Allen, A. Fakorede, A. C. R. Martin, The SAAPdb web resource: A large-scale structural analysis of mutant proteins, Hum Mutat 30 (2009) 616–624.

[22] N. S. Al-Numair, A. C. R. Martin, The SAAP pipeline and database: Tools to analyze the impact and predict the pathogenicity of mutations, BMC Genomics 14 (2013) 1–11.

[23] E. Capriotti, R. Calabrese, R. Casadio, Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information, Bioinformatics 22 (2006) 2729–2734.

[24] E. Capriotti, R. Calabrese, P. Fariselli, P. L. Martelli, R. B. Altman, R. Casadio, WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation, BMC Genomics 14(Suppl 3) (2013) S6.

[25] J. Tian, N. Wu, X. Guo, J. Guo, J. Zhang, Y. Fan, Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines, BMC Bioinformatics 8 (2007) 450–464.

[26] L. Bao, M. Zhou, Y. Cui, nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms, Nucleic Acids Res 33 (2005) W480–W482.

[27] B. Li, V. G. Krishnan, M. E. Mort, F. Xin, K. K. Kamati, D. N. Cooper, S. D. Mooney, P. Radivojac, Automated inference of molecular mechanisms of disease from amino acid substitutions, Bioinformatics 25 (2009) 2744–2750.

[28] V. Acharya, H. A. Nagarajaram, Hansa. An automated method for discriminating disease and neutral human nsSNPs, Human Mutation 2 (2012) 332–337.

[29] J. M. Schwarz, C. Rödelsperger, M. Schuelke, D. Seelow, MutationTaster evaluates disease-causing potential of sequence alterations, Nature Methods 7 (2010) 575–576.

[30] A. González-Pérez, N. López-Bigas, Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel, Am J Hum Genet 88 (2011) 440–449.

[31] J. Bendl, J. Stourac, O. Salanda, A. Pavelka, E. Wieben, J. Zendulka, J. Brezovsky, J. Damborsky, PredictSNP: Robust and accurate consensus classifier for prediction of disease-related mutations, PLOS Computational Biology 10 (2014) e1003440.

[32] D. E. V. Pires, D. B. Ascher, T. L. Blundell, DUET: a server for predicting effects of mutations on protein stability via an integrated computational approach, Nucleic Acids Research 42(W1) (2014) W314–W319.

[33] L. F. Stead, I. C. Wood, D. R. Westhead, Kvsnp: Accurately predicting the effect of genetic variants in voltage-gated potassium channels, Bioinformatics 27 (2011) 2181–2186.

[34] L. E. M. McMillan, A. C. R. Martin, Automatically extracting functionally equivalent proteins from SwissProt, BMC Bioinformatics 9 (2008) 418–418.