Using structural bioinformatics to investigate the impact of non synonymous SNPs and disease mutations: scope and limitations

Joke Reumers[1], Joost Schymkowitz[1] and Fréderic Rousseau[*1]

[1]Switch Laboratory, VIB, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

Email: Joke Reumers - joke.reumers@vub.ac.be; Joost Schymkowitz - joost.schymkowitz@vub.ac.be; Fréderic Rousseau*- frederic.rousseau@vub.ac.be;

*Corresponding author

Abstract

Background: Linking structural effects of mutations to functional outcomes is a major issue in structural bioinformatics, and many tools and studies have shown that specific structural properties such as stability and residue burial can be used to distinguish neutral variations and disease associated mutations.

Results: We have investigated 39 structural properties on a set of SNPs and disease mutations from the Uniprot Knowledge Base that could be mapped on high quality crystal structures and show that none of these properties can be used as a sole classification criterion to separate the two data sets. Furthermore, we have reviewed the annotation process from mutation to result and identified the liabilities in each step.

Conclusions: Although excellent annotation results of various research groups underline the great potential of using structural bioinformatics to investigate the mechanisms underlying disease, the interpretation of such annotations cannot always be extrapolated to proteome wide variation studies. Difficulties for large-scale studies can be found both on the technical level, i.e. the scarcity of data and the incompleteness of the structural tool suites, and on the conceptual level, i.e. the correct interpretation of the results in a cellular context.

Background

The molecular phenotype of a coding non synonymous SNP or disease associated mutation describes the functional and structural properties of a protein that are affected by a single amino acid substitution [25]. In this study we want to address whether the concept of the *in silico* determined molecular phenotype can be employed for large-scale classification of SNPs and disease mutations. The attempt to classify a large set of mutations based on an incomplete molecular phenotype may seem naive at first glance, had it not been suggested that individual properties such as protein stability, the accessibility of the amino acid substitution site, and the location of variants in surface pockets are predictive determinants of the phenotypic effect of a variation [1–4]. A comparative study of protein stability predictors by Blundell and co-workers demonstrated that although protein stability changes caused by mutation can be relatively accurately estimated *in silico*, these

predictions by themselves do not yield accuracy on large-scale classification between benign and disruptive mutations [5–7].

Furthermore, computational analyses rely heavily on the quality of the data under scrutiny and the computational methods used to evaluate these data. Before investigating 39 structural properties of proteins and amino acid substitutions for their predictive power regarding SNP classification, we have investigated what major liabilities are encountered when implementing an structural approach to SNP annotation and classification. The results are compared with those achieved by the best performers among the state-of-the-art tools.

Results and Discussion

In this study we have identified the common issues that are encountered when performing large-scale analyses of structural properties of human coding variation. The first issue concerns the availability of structural data for nsSNPs and disease mutations, while the second involves the availability of computational tools to predict structural properties. The last issue concerns the quality of classification: are the training and evaluation data sets used in the analyses sufficient to extrapolate results for larger studies, and do the properties used have sufficient predictive power to separate the two data sets?

Structural coverage of human genetic variation

Despite structural genomics projects, the gap between sequence and structural information is still wide, and the coverage of variation data with structural data is estimated to be as low as 14% [4]. We have investigated the boundaries of structural coverage by varying the quality requirements on the structural model (Supplementary Figure S1A), the sequence identity between query sequence and modelled structure (Figure S1B), the percentage of the wild type sequence covered by the structural model (Figure S1C), and the length of the alignment between query and target (Figure S1D). Without applying any restrictions, about 12% of all nsSNPs present in the Ensembl Variation Database (release 44) can be mapped on a structural model, in accordance with the estimate cited previously. However, this percentage is valid only when no restrictions regarding sequence identity, sequence coverage or structure quality are applied. Our standard restrictions on building high-confidence structural models using the FoldX force field are X-ray structures with a resolution lower than 2.5 Å and sequence identity higher than 80%. Applying these restrictions to the Ensembl data results in a data set of 5416 nsSNPs (circa 4% of the data, Figure S1B).

Predictability of structural properties

The second issue for a large-scale structural bioinformatics approach is the structural properties that are predictable with state of the art tools: how well can we describe the structural behaviour of a protein and its mutants? Previous structural studies have identified protein stability, aggregation and misfolding as determinants of correct functioning on the single protein level [7,11,12]. Mutations affecting the functional sites of a protein, such as DNA, ligand and protein interaction sites, are not considered within this scope, but the investigation of these sites will most certainly be of great importance to assess the impact of amino acid substitutions.

Tools have been developed that describe the structure and dynamics of a protein: stability, aggregation, amyloidosis, and folding. We have used computational methods that are capable of assessing the effects of a mutation on protein stability (FoldX), aggregation (Tango) and amyloidosis (Waltz). Although algorithms exist that can predict folding of small single domain proteins (e.g. Rosetta [13], FoldX [14], SimFold [15]), to date no computational method exists that can predict folding events on large multi-domain proteins, or that is applicable in genome wide studies.

Although we have not investigated protein-protein interactions in this study, we have included an analysis of the binding of proteins to molecular chaperones, as it is directly related to correct folding of the protein. The high abundance of chaperones in the cell emphasises their crucial role in the cell [16], but this is not reflected in the availability of computational tools for chaperone binding. We have used the only available tool, the Hsp70 binding predictor Limbo [17], to assess chaperone binding variation caused by amino acid alteration.

The predictive power of structural properties

Following the recommendations of Care *et al* [18], we have used the SwissProt annotated disease and

2

polymorphism data (SwissProt Variation Index release 52) as the evaluation data for our analyses. Mapping of these variants on high quality structural models (X-ray structures with resolution ≤ 2.5Å, sequence identity with the model above 80%) yielded a data set of 240 positive (disease-associated) mutations and 400 negative variations (neutral nsSNPs) in 98 proteins. To ensure that the analyses are comparable, we applied the sequence based predictors to the same small data set as the predictors that use 3D structures or structural models.

Before we evaluated the discriminative power of the individual structural parameters, we wanted to assess whether our data showed distinguishable patterns for three important parameters. The first two criteria, stability difference and the degree of burial of the mutation site, have previously been identified as providing information about the severity of a mutation [4, 19]. The third criterion is difference in aggregation propensity, which has been cited as likely to be an important factor in disease susceptibility [12, 20] but thus far has not been applied in a proteome wide mutation analysis.

Figure 1 shows the distributions for the stability differences (A) and differences in aggregation propensity (B) between wild type and variant proteins, and the burial of the mutation site (C). The first observation of both the stability and the aggregation analysis is that the observed changes are not discrete but follow a smooth distribution from negative to positive change. Second, there are noticeable differences between SNPs and disease mutations, but they cannot be distinguished by a simple cut-off value on the output, as there is large overlap between the distributions. This is confirmed by the P-values obtained from paired student t-tests, which are 0.96 for the stability distributions, 0.99 for the aggregation distributions, and 0.99 for the burial distributions, respectively. For the stability distributions, we see that disease mutations are generally more destabilising than SNPs, but their distributions overlap largely. A similar analysis has been performed on SwissProt variants using the Site Directed Mutator stability predictor [7], and the distributions of stability differences of disease mutations and neutral variations are similar to our findings.

In a first series of properties to test as classifiers, we have investigated 15 properties of the amino acid substitution site that contribute to the assessment of the effect of the mutation using the FoldX algorithm (Table 3). Cut off values were generated that var-

ied between the minimal and maximal values measure for the specific property, and the true and false positive rate, and the Matthews correlation coefficient (MCC) were calculated for each cut-off value. Table 3 lists the data for both the best MCC and the MCC90, i.e. the coefficient that is measured at high specificity (true negative rate = 90%). The corresponding ROC curves for these analyses can be found in Supplementary Figure S1.

The same strategy was then applied to predicted values of structural differences between mutant and wild type proteins (24 properties). Statistics were calculated for stability and entropy parameters, as well as for differences concerning protein aggregation, amyloidosis and chaperone binding (Table 4, Supplementary Figure S2).

The results obtained from these detailed analyses are unanimous: none of the parameters evaluated can be used to separate the data. All MCC values are close to zero, and thus the predictions are no better than a random predictor would perform on the data. The high accuracy of FoldX for stability estimation has been proven in various studies [6,9,10], so we have high confidence in our stability estimations. In accordance with the analyses of [7], we find that high stability differences alone are no sufficient criterion to distinguish deleterious mutations and neutral variation. These results show that the dominant effect of for instance stability that was proposed in earlier large-scale studies [4, 22] can not be always generalised for other data.

The fact that none of the properties representing conformational differences between wild type and variant protein contain enough information to distinguish neutral and deleterious variation implies that large-scale classification based on singular structural properties is not feasible and requires a better understanding of how the complex interplay between biophysical and biochemical properties of a protein conspire to different tolerance for mutations in different proteins.

Recent studies that combine structural and evolutionary information using machine learning techniques are able to classify relatively large data sets obtained for the SwissProt database successfully (summarised in Table S2). Machine learning approaches suggest that data integration is indeed the way forward, but the creation of this black box style of classifier does not offer insight into the biological processes. In the same way that using evolutionary information to classify SNPs obscures the how and

why a specific mutation is deleterious, using black box machine learning methods will not teach us what the underlying reason of disease is. Although knowing that an amino acid is critical for correct function is of course useful, in a structural bioinformatics approach the focus is more on the molecular mechanism underlying disease.

A simple combination of the SNPeffect structural bioinformatics toolsuite on our evaluation data set showed that in our case, at least a linear combination of these methods is not sufficient to classify the data (TPR = 0.73, TNR= 0.27, MCC=0). A large part of the polymorphism data is predicted to have deleterious effect. To assess the "predictiveness" of our data set, we applied the well-established evolutionary method SIFT [24] to our data and found that SIFT was also not able to classify effectively. In fact the results were even worse than our naive classifier (TPR=0.69, TNR=0.21, MCC=-0.12).

As an illustration of the influence of the data set used for evaluation on the performance of a predictor, we list the results for the variation in performance of SNP classification of SIFT, that uses evolutionary information to label SNPs (Supplementary Table S3). The Matthews correlation coefficient varies between -0.12 on our data set over 0.25 on human mutagenesis data, up to 0.59 on the HIV-1 protease mutagenesis set in the original SIFT paper [24]. This is yet another informative example on how crucial the choice of training and test data are to build and evaluate predictors: generalisation of results is only possible when the training data are expressive enough to represent the entire feature space.

Conclusions

The concept of using the molecular phenotypic effect of a nsSNP to assess its effect on the structure and function of the protein it alters was first introduced by Bork and co-workers [25]. The question has been raised to how much of this molecular phenotype is necessary to evaluate the contribution of a SNP to a disease phenotype: are there singular dominant properties that determine the impairment of structure and function, or do we need to consider the full ensemble of molecular properties to interpret the impact of the SNP? Other research groups have proposed that single properties such as stability [4] and solvent accessibility [1] can be used to classify SNPs. We have examined all the individual struc-

tural bioinformatics tools that were proposed in the SNPeffect toolsuite [26] for their ability to act as a binary classifier for deleterious and neutral SNPs. Neither of the individual properties that were examined could serve this purpose. Because several approaches were able to classify similar data sets as the one we have used, we applied the most used evolutionary method, SIFT [23], to our data set. As it was not able to classify our data set accurately, we argued that generalisation of the results presented by the state of the art classifiers might be an important issue. We illustrated this problem with the variability of performance of SIFT on 8 different data sets used in various analyses.

From these analyses we concluded that strict classification of SNPs is not feasible at the time, both because there are still many technical difficulties to overcome, and because the biological interpretation of the molecular phenotype in relation to a disease phenotype is a complex matter. Even at the single molecule level, we cannot assess how tolerant a specific protein is to structural variation. The inherent rigidity of a protein might influence the change in stability that is allowed before severe conformational changes are introduced. Furthermore, on the cellular level biological interpretation is even harder: we can not predict the role of the protein quality control system plays in this tolerance level, not all interactions are described at the molecular level, and much more. Even if we can predict the molecular effect accurately, this might not necessarily result in a disease phenotype because of functional redundancy of the protein.

However, not being able to classify human variation into disease mutations and neutral or beneficial variation does not mean that this approach or the methods developed are useless. By using high quality bioinformatics tools, we can select from a large pool of variations the candidates that are interesting for detailed investigation. This in itself is a valuable contribution, because the amount of variation data available is too massive to be investigated experimentally. *In silico* analyses can and will be used successfully as an addition to *in vitro* and *in vivo* studies.

4

Methods

Assembly of data sets

Statistics on the structural coverage and validation status of human non synonymous coding SNPs were performed on data from the Ensembl human variation database release 44, containing 12.2 million SNPs, of which 133698 cause an amino acid variation in a known transcript. The mapping of SNPs on protein structures was evaluated using the "ensppdbmapping" DAS service provided by the SPICE server [27]. Positive and negative data sets for the evaluation of SNP classification were designed with data from the SwissProt variation index [28] in the UniProt knowledge base (version 52.0, March 2007, [29]) that were mapped onto known PDB structures and high quality homologs thereof. The quality criteria described in the results section (models with resolution of 3 Åor higher, sequence identity of 80% or more) lead to structural models of 400 SNPs (negative) and 240 disease associated mutations (positive).

Structural bioinformatics tools

We have used the FoldX force field [33] for all mutant properties regarding structural location, protein stability and its various components, the Tango [34] and Waltz [35, submitted] algorithms to assess the propensity for aggregation of wild type and variant proteins, and the Limbo algorithm [17, submitted] to evaluate the chaperone-binding properties of amino acid sequences. A novel tool developed by Lenaerts *et al* (unpublished) was used to estimate the entropy of a specific amino acid site in a high-resolution structure. Detailed descriptions of these five tools can be found in the Supplementary Material.

Authors contributions

Conceived and designed the experiments: JR JS FR. Performed the experiments: JR. Analysed the data: JR JS FR. Wrote the paper: JR.

Acknowledgements

References

1. Chasman D, Adams RM: **Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: Structure-based assessment of amino acid variation**. *J Mol Biol* 2001, **307**(2):683–706.

2. Ferrer-Costa C, Orozco M, de la Cruz X: **Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties**. *J Mol Biol* 2002, **315**(4):771–786.

3. Stitziel NO, Tseng YY, Pervouchine D, Goddeau D, Kasif S, Liang J: **Structural location of disease-associated single-nucleotide polymorphisms**. *J Mol Biol* 2003, **327**(5):1021–1030.

4. Yue P, Li Z, Moult J: **Loss of protein structure stability as a major causative factor in monogenic disease**. *J Mol Biol* 2005, **353**(2):459–473.

5. Worth CL, Burke DF, Blundell TL: **Estimating the effects of single nucleotide polymorphisms on protein structure: how good are we at identifying likely disease associated mutations?** In *Proceedings of Molecular Interactions - Bringing Chemistry to Life* 2006.

6. Burke DF, Worth CL, Priego EM, Cheng T, Smink LJ, Todd JA, Blundell TL: **Genome bioinformatic analysis of nonsynonymous SNPs.** *BMC Bioinformatics* 2007, **8**:301.

7. Worth CL, Bickerton GRJ, Schreyer A, Forman JR, Cheng TMK, Lee S, Gong S, Burke DF, Blundell TL: **A structural bioinformatics approach to the analysis of nonsynonymous single nucleotide polymorphisms (nsSNPs) and their relation to disease.** *J Bioinform Comput Biol* 2007, **5**(6):1297–1318.

8. Boutselakis H, Dimitropoulos D, Fillon J, Golovin A, Henrick K, Hussain A, Ionides J, John M, Keller PA, Krissinel E, McNeil P, Naim A, Newman R, Oldfield T, Pineda J, Rachedi A, Copeland J, Sitnov A, Sobhany S, Suarez-Uruena A, Swaminathan J, Tagari M, Tate J, Tromm S, Velankar S, Vranken W: **E-MSD: the European Bioinformatics Institute Macromolecular Structure Database.** *Nucleic Acids Res* 2003, **31**:458–462.

9. Guerois R, Nielsen JE, Serrano L: **Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations**. *J Mol Biol* 2002, **320**(2):369–387.

10. Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS: **The stability effects of protein mutations appear to be universally distributed**. *J Mol Biol* 2007, **369**(5):1318–1332.

11. Steward RE, MacArthur MW, Laskowski RA, Thornton JM: **Molecular basis of inherited diseases: a structural perspective.** *Trends Genet* 2003, **19**(9):505–513.

12. DePristo M, Weinreich D, Hartl D: **Missense meanderings in sequence space: A biophysical view of protein evolution.** *Nature Reviews Genetics* 2005, **AOP**.

13. Simons KT, Bonneau R, Ruczinski I, Baker D: **Ab initio protein structure prediction of CASP III targets using ROSETTA.** *Proteins* 1999, **Suppl 3**:171–176.

5

14. Serrano L, Guerois R: **Fold-X: An algorithm to predict and engineer folding pathways.** *Abstr Pap Am Chem Soc* 2001, **221**:U395–U395.

15. Fujitsuka Y, Chikenji G, Takada S: **SimFold energy function for de novo protein structure prediction: consensus with Rosetta.** *Proteins* 2006, **62**(2):381–398.

16. Soti C, Csermely P: **Protein stress and stress proteins: implications in aging and disease.** *J Biosci* 2007, **32**(3):511–515.

17. Van Durme J, Maurer-Stroh S, Wilkinson H, Rousseau F, Schymkowitz J: **Accurate prediction of the sequence determinants of DnaK-peptide binding via a method that integrates homology modelling and experimental data**. *Submitted* 2007.

18. Care MA, Needham CJ, Bulpitt AJ, Westhead DR: **Deleterious SNP prediction: be mindful of your training data!** *Bioinformatics* 2007, **23**(6):664–672.

19. Ramensky V, Bork P, Sunyaev S: **Human non-synonymous SNPs: server and survey**. *Nucleic Acid Res* 2002, **30**(17):3894–3900.

20. Worth CL, Blundell TL: **Estimating the effects of SNPs on protein structure: loss of protein interactions and stability as indicators of mis-function and disease-association.** *Curr Top Biochem Res* 2008, **In press**.

21. Stitziel NO, Binkowski TA, Tseng YY, Kasif S, Liang J: **TopoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association**. *Nucleic Acid Res* 2004, **32**:D520–D522.

22. Yue P, Melamud E, Moult J: **SNPs3D: candidate gene and SNP selection for association studies.** *BMC Bioinformatics* 2006, **7**:166.

23. Ng PC, Henikoff S: **SIFT: predicting amino acid changes that affect protein function**. *Nucleic Acid Res* 2003, **31**(13):3812–3814.

24. Ng PC, Henikoff S: **Predicting deleterious amino acid substitutions**. *Genome Res* 2001, **11**(5):863–874.

25. Sunyaev S, Lathe Wr, Bork P: **Integration of genome data and protein structures: prediction of protein folds, protein interactions and "molecular phenotypes" of single nucleotide polymorphisms.** *Curr Opin Struct Biol* 2001, **11**:125–130.

26. Reumers J, Conde L, Medina I, Maurer-Stroh S, Van Durme J, Dopazo J, Rousseau F, Schymkowitz J: **Joint annotation of coding and non-coding single nucleotide polymorphisms and mutations in the SNPeffect and PupaSuite databases.** *Nucleic Acids Res* 2008, **36**(Database issue):D825–9.

27. Prlic A, Down TA, Hubbard TJ: **Adding some SPICE to DAS**. *Bioinformatics* 2005, **21 Suppl 2**:ii40–1.

28. Yip YL, Famiglietti M, Gos A, Duek PD, David FPA, Gateau A, Bairoch A: **Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase.** *Hum Mutat* 2008, **29**(3):361–366.

29. UniProt Consortium: **The Universal Protein Resource (UniProt)**. *Nucleic Acids Res* 2007, **35**(Database issue):D193–7.

30. Zweig MH, Campbell G: **Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine.** *Clin Chem* 1993, **39**(4):561–577.

31. Matthews BW: **Comparison of the predicted and observed secondary structure of T4 phage lysozyme**. *Biochim Biophys Acta* 1975, **405**(2):442–451.

32. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H: **Assessing the accuracy of prediction algorithms for classification: an overview.** *Bioinformatics* 2000, **16**(5):412–424.

33. Schymkowitz JWH, Rousseau F, Martins IC, Ferkinghoff-Borg J, Stricher F, Serrano L: **Prediction of water and metal binding sites and their affinities by using the Fold-X force field**. *Proc Natl Acad Sci USA* 2005, **102**(29):10147–10152.

34. Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L: **Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins**. *Nat Biotechnol* 2004, **22**(10):1302–1306.

35. Maurer-Stroh S, Kuemmerer N, Lopez de la Paz M, Martins I, Reumers J, Serrano L, Rousseau F, Schymkowitz J: **Accurate prediction of sequence determinants of amyloid formation using the Waltz algorithm**. *Submitted* 2007.

Figures

Figure 1 - Distributions for the major structural criteria in the disease and polymorphism datasets.

White = disease mutations, grey = polymorphisms. **A.** Stability difference as calculated by the FoldX force field (in kcal.mol^{-1}). **B.** Difference in aggregation propensity as calculated by the Tango algorithm. Values close to neutral changes (in the range $[-50, 50]$) are left out for display purposes. **C.** Distribution of degree of burial of the amino acid substitution site.

Tables

Table 1 - Summary of structural coverage of SNP data.
Several criteria resulting from the above analyses are applied to assess the structural coverage and reliability of that coverage of human SNPs in the Ensembl database, as well as the overlap of the structural coverage with quality parameters for the validation and frequency status of the polymorphism data.

Properties	# SNPs	% SNPs
nsSNPs covered by high quality structural data		
No additional criteria	9877	7.4
Sequence coverage>80 or alignment length> 100	8238	6.2
Sequence identity>80	5416	4.1
Sequence coverage>80 or alignment length> 100, and sequence identity>80	5318	4.0
Highly reliable nsSNPs covered by high quality structural data		
Doublehit validation status, MAF>0.01	680	0.51
Doublehit validation status, MAF>0.01, sequence identity>80	229	0.17
Doublehit validation status, MAF>0.01, sequence coverage>80 or alignment length> 100	446	0.33
Doublehit validation status, MAF>0.01, sequence coverage>80 or alignment length> 100, and sequence identity>80	209	0.16

Table 2 - Predictive power of structural properties of the modeled variant proteins.
FoldX was used to evaluate both the overall stability contribution of the amino acid substitution site in the modeled structure and the various factors involved in this stability. The entropy of the variant amino acid was calculated using a sampling strategy to assess the possible side chain conformations allowed at the substitution site. Both stability and entropy were calculated for all mutations and for a subset of buried mutations (side chain burial < 0.5) and surface mutations (side chain burial ≥ 0.5). Corresponding ROC curves are shown in Supplementary Figure S2.

Table 1					
Property	**FPR**	**TPR**	**Best MCC**	**Threshold**	**MCC90**
FoldX energy evaluation					
Overall stability of residue	14	33	0.22	1.61	0.19
Backbone H bond	32	72	0.40	-1.05	0.22
Sidechain H bond	99	100	0.07	-1.76	<0
Electrostatics	86	93	0.11	-0.10	-0.01
Entropy side chain	59	80	0.22	0.32	0.05
Entropy main chain	13	27	0.18	1.96	0.10
Van der Waals contribution	25	47	0.23	-0.98	0.15
Solvation hydrophobic	10	22	0.16	-0.6	0.16
Solvation polar	42	70	0.28	1.5	0.06
Van der Waals clash	18	33	0.17	0.22	0.15
Side chain burial	51	67	0.16	0.43	-0.1
Main chain burial	59	83	0.26	0.73	0.05
Entropy by sampling of possible side chain conformations					
Entropy side chain	72	84	0.15	0.93	0

Table 3 - Predictive power of the differences between wild type and variant proteins for different structural properties.

FoldX was used to evaluate both the overall stability difference between wild type and variant structure, and the constituting contributions leading to this stability difference. The entropy difference caused by the amino acid substitution was calculated using a sampling strategy to assess the possible side chain conformations allowed at the substitution site. Both stability and entropy difference were calculated for all mutations and for a subset of buried mutations (side chain burial < 0.5) and surface mutations (side chain burial $\geq$ 0.5). Corresponding ROC curves are shown in Supplementary Figure S3.

Property	FPR	TPR	Best MCC	Threshold	MCC90
FoldX energy evaluation					
Overall stability difference	73	85	0.15	-0.45	0.14
Overall stability diff. (surface)	0	8	0.2	3.1	0.13
Overall stability diff. (buried)	21	44	0.25	2.64	0.12
Backbone clash	91	99	0.18	-1.00	-0.02
Backbone H bond	59	83	0.26	-0.025	0.06
Sidechain H bond	79	92	0.18	-0.13	-0.14
Electrostatics	6	18	0.18	0.15	0.16
Entropy main chain	6	18	0.18	0.15	0.04
Entropy side chain	64	74	0.11	-0.125	-0.05
Solvation hydrophobic	57	75	0.19	-0.15	-0.03
Solvation polar	22	36	0.15	0.20	-0.05
Torsion clash	1	3	0.07	1.00	-0.05
Van der Waals contribution	7	14	0.11	0.89	0.10
Van der Waals clash	98	100	0.10	-1.60	0.02
Entropy difference by sampling of possible side chain conformations					
FoldX entropy difference	85	92	0.11	-1.85	-0.02
FoldX entropy diff. (buried)	96	100	0.14	-2.70	-0.05
FoldX entropy diff. (surface)	37	57	0.20	-0.10	0.02
Aggregation properties					
Tango	1	3	0.07	39.9	0
Tango (positive, more aggr.)	14	22	0.10	16.37	0
Tango (negative, less aggr.)	69	78	0.10	-8.00	0
Waltz	0	1	0.07	748.97	0
Waltz (positive, more aggr.)	16	21	0.06	677.15	0
Waltz (negative, less aggr.)	99	100	0.07	-2412.78	0
Limbo	17	33	0.18	5.45	0

Additional Files

Figure 1 – figure1.pdf

Additional file 2 — supplementary.pdf

Several of the less critical figures and tables are added as supplementary material, together with detailed descriptions of the structural bioinformatics tools used.

Using structural bioinformatics to investigate the impact of non synonymous SNPs and disease mutations: scope and limitations
Supplementary Material

Joke Reumers[1], Joost Schymkowitz[1] and Fréderic Rousseau[*1]

[1]Switch Laboratory, VIB, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

Email: Joke Reumers - joke.reumers@vub.ac.be; Joost Schymkowitz - joost.schymkowitz@vub.ac.be; Fréderic Rousseau[*]- frederic.rousseau@vub.ac.be;

[*]Corresponding author

Methods

Structural bioinformatics tools

FoldX

The FoldX force field was developed for the fast and accurate estimation of the free change upon mutation on the stability of a protein or a protein complex [23–26]. It uses an all-atom representation of these macromolecules, and has been validated on a test database of more than 1000 mutants from more than 20 different proteins. It currently yields a correlation of 0.78 with a standard deviation of 0.41 kcal/mol.

Modelling and evaluation of mutations in FoldX is performed with the *BuildModel* command. It is used first to model a homologous sequence on a structural model and to optimise the side chains to fit the new sequence, and then to evaluate the effect of a single amino acid variation. The Gibbs free energy of a protein is calculated with the Stability command. The various structural parameters used in the classification tests (backbone clash, backbone H bond formation , sidechain H bond formation, electrostatics , solvation of hydrophobic residues, solvation of polar residues, torsion clash, Van der Waals contribution,Van der Waals clash)

Entropy calculations based on side chain sampling

In addition to the entropy calculations intrinsic to the FoldX force field, we use a novel method based on extensive sampling of side chain conformations as developed by Lenaerts *et al.* (unpublished). The sampling method produces for each side chain the probability ($P(X)$) of finding the residue's side chain in a particular conformational state. From these probabilities entropy can easily be derived:

$$H(X) = -\sum iP(x_i)log_2P(x_i)$$

The method uses a rotamer database based on conditional statistics of dihedral angles derived from the WHAT IF data set [27]. All amino acids from this data and their corresponding dihedral angles ($10°$ bin) were used to derive the following probabilities: $P(\chi_i), P(\chi_i|\chi_{i-1})$ and $P(\chi_i|\chi_{i-1}, \chi_{i-2})$, except for $\chi_1(P(\chi_1)$ and $P(\chi_1|\phi, \psi))$. A set of n random rotamers can be derived from the probability distribution thus calculated. This will allow sampling of rotamers with greater resolution than classical rotamer libraries.

The sampling itself is performed by Monte Carlo based sampling method with Metropolis criterion (at 298K). The Metropolis criterion states that a certain conformational change is accepted with a probability p that depends on the free energy change $\Delta\Delta G$ associated with the conformational change as given

by the following formula:

$$p = 1 \, if \, \Delta\Delta G < 0$$

$$p = e^{-\frac{\Delta\Delta G}{RT}} \, if \, \Delta\Delta G \geq 0$$

The free energy of each change is determined with FoldX.

Tango

The β-aggregation prediction algorithm Tango [28] uses a statistical mechanics approach to represent a competition between major conformational states: the random coil and the native conformations, as well as β-turn, α-helix and β-aggregate. Two windows of variable length slide over the sequence, and each such window can populate these conformational states according to a Boltzmann distribution. The frequency of population of each structural state for a given segment will be relative to its energy, which is derived from statistical and empirical parameters. To predict the β-aggregating segments of a peptide, Tango calculates the partition function of the phase space involving these conformational states. In our analysis we have used Tango to calculate the difference in aggregation tendency that results from an single amino acid variation.

Waltz

Current methods for the prediction of the sequence determinants of amyloidosis suffer from two major problems: overpredicting amorphous cross β aggregates and missing amylogenic sequences that are enriched in the polar Q and N residues, such as the prion protein. The Waltz algorithm [29, submitted] tackles these problems by taking into account amyloid hexapeptides from 48 new amyloid forming sequences, derived from 31 proteins. About half the proteins in this extended data set were not previously known to contain amyloidogenic sequences such as presenilin-2, titin and myosin. Waltz combines terms from amino acid sequence scoring in the learning set, physical property analysis and homology modelling. The method shows 84% sensitivity at 92% specificity on the AmylHex data set [30], and correctly identifies mutations in human proteins known to be associated with amyloid deposition.

Limbo

Limbo is a Hsp70 binding site predictor that was built using a dual method combining sequence and structural information [31, submitted]. Experimental DnaK binding data of 53 non-redundant peptide sequences was used to generate a sequence-based position-specific scoring matrix (PSSM) based on logarithm of the odds scores. Following an *in silico* alanine scan of the substrate peptide in the crystal structure of a DnaK-substrate complex (PDBID 1DKX Zhu1996) using FoldX, a structure-based PSSM that reflects the individual contribution of certain substrate residue types for DnaK binding was generated. The Limbo DnaK binding site predictor was obtained by combining the structure-based PSSM with a normalisation factor of 0.2 with the sequence-based PSSM. Limbo is able to correctly predict 89% of the true positives in a tested peptide set (high sensitivity), with a concurrent amount of only 5.9% false positives for a specific score threshold (high specificity). The robustness of the predictor was evaluated with a cross-validation test, resulting in a true positive rate of 72% true positives and a false positive rate of 5.9%. The predictor was able to identify an entire known DnaK binding site in the heat-shock promoter σ 32 [32]. We have used Limbo to rank mutated proteins according to their DnaK binding affinity.

2

Tables

Supplementary Table S1 - Types of data sets used to train and test SNP classifiers.

Origin data set	Size of data set	Number of studies	References
Neutral variations			
Mutagenesis studies	111-3706	9	[1–9]
Orthologs	888-16682	3	[3, 9, 10]
SwissProt SNP	502-12944	6	[3, 8, 11–14]
OMIM	558	1	[15]
dbSNP	5177-21471	2	[16, 17]
Disease mutations			
Mutagenesis studies	159-1750	8	[1–9]
COSMIC database	879	1	[18]
HGMD	3768-10263	1	[9]
OMIM	879-2249	5	[3, 8, 13, 15, 18]
SwissProt Disease	175-9610	9	[3, 8, 10–14, 19, 20]
Data [21]	209	1	[20]
Data [22]	185	2	[19, 20]

Supplementary Table S2 - Performance of state-of-the-art predictors on representative data sets.
The performance of a few selected tools on SwissProt disease associated mutations and SNP data are shown.

Study	Method	FPR	FNR	TPR	TNR	MCC	Size set
Bao *et al* [11]	Random Forest	0.3	0.24	0.76	0.7	0.46	205
Capriotti *et al* [13]	HybridMeth	-	-	-	-	0.46	21185
Karchin *et al* [14]	SVM	0.2	0.19	0.81	0.8	0.61	3691
Ng & Henikoff [19]	SIFT	0.19	0.31	0.69	0.81	0.50	5333
Wang & Moult [20]	Stability	0.3	0.1	0.9	0.7	0.61	262
Worth *et al* [16]	Combined	0.09	0.68	0.32	0.91	0.28	9143
Yue & Moult [9]	SVM	0.15	0.26	0.74	0.85	0.59	6077

Supplementary Table S3 - Variation of the performance of SIFT on different data sets.

Study	Dataset	FPR	FNR	TPR	TNR	MCC
Bao *et al* [11]	Test set	0.33	0.38	0.62	0.67	0.29
Saunders *et al* [8]	Human	0.4	0.35	0.65	0.6	0.25
Ng & Henikoff [7]	lac I repressor	0.22	0.43	0.57	0.78	0.36
Ng & Henikoff [7]	HIV 1-protease	0.3	0.12	0.88	0.7	0.59
Ng & Henikoff [7]	T4 lysozyme	0.41	0.28	0.72	0.59	0.31
Ng & Henikoff [19]	SwissProt disease	0.19	0.31	0.69	0.81	0.50
Worth *et al* [16]	SwissProt + dbSNP	0.41	0.29	0.71	0.59	0.30
Our evaluation	SwissProt	0.79	0.31	0.69	0.21	-0.12

Figures
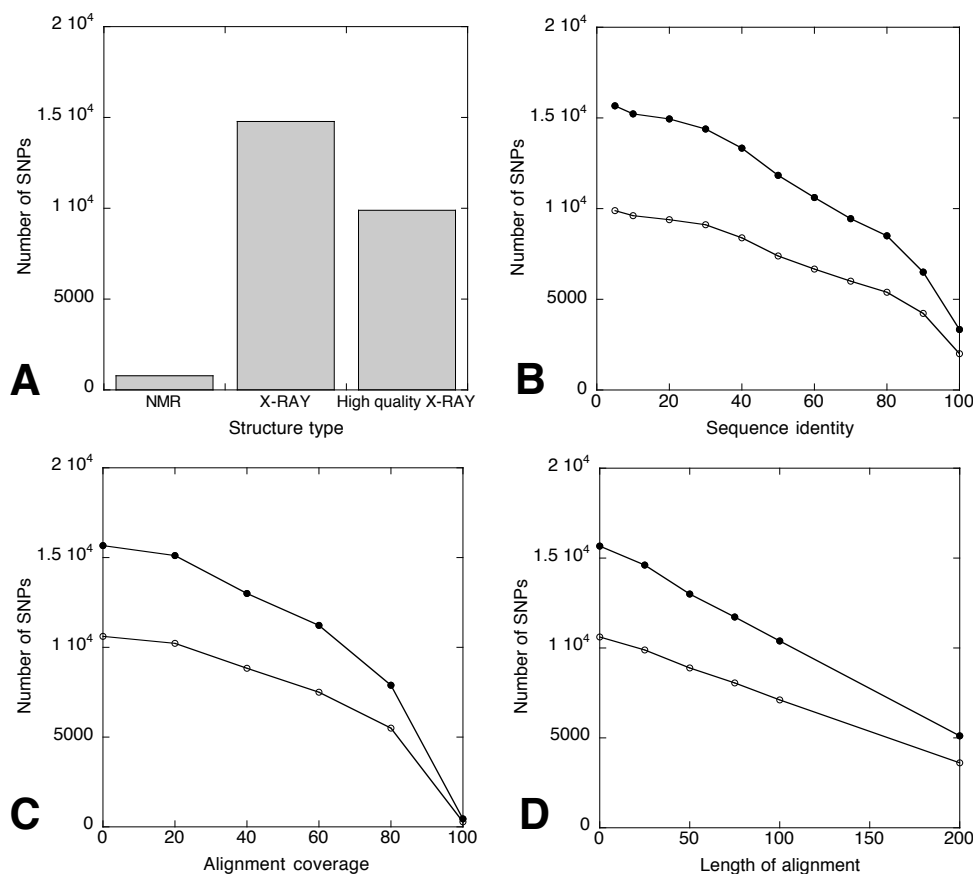
Figure S1. Structural coverage of Ensembl non synonymous SNP data. A. Number of SNPs in structures determined by NMR and X-ray crystallography studies or models of these structures. 11% of all non synonymous SNPs can be mapped on crystallography structures, and 7% of all SNPs can be modeled on a high-quality X-ray structure (resolution ≤ 2.5Å). **B. Number of SNPs covered by structural data versus the sequence identity between the query sequence and the structural model.** The number of SNPs that can be modeled on X-ray structures (•) decreases from 15% of all nsSNPs (15685 nsSNPs, 5% sequence identity) to 2.5% (3341) of all SNPs for which the structure of the wild type sequence has been determined experimentally (100% sequence identity). When only high quality structures are considered (∘), this amount is reduced by half to 7.4% for a sequence identity of 5% and 1.5% for exact models. **C. Number of SNPs covered by structural data versus the sequence coverage of the wild type sequence.** There are almost no SNPs for which the full length of the protein sequence is covered (100% coverage), but for 80% coverage almost 8000 SNPs can be selected, of which circa 5500 in high quality structures. **D. Number of SNPs covered by structural data versus the length of the alignment between protein sequence and structural model.** About a third of the SNPs that can be modeled are located in a structural alignment that is less than 100 amino acids long, both for models based on all X-ray structures (•) and based on high resolution structures only (∘).
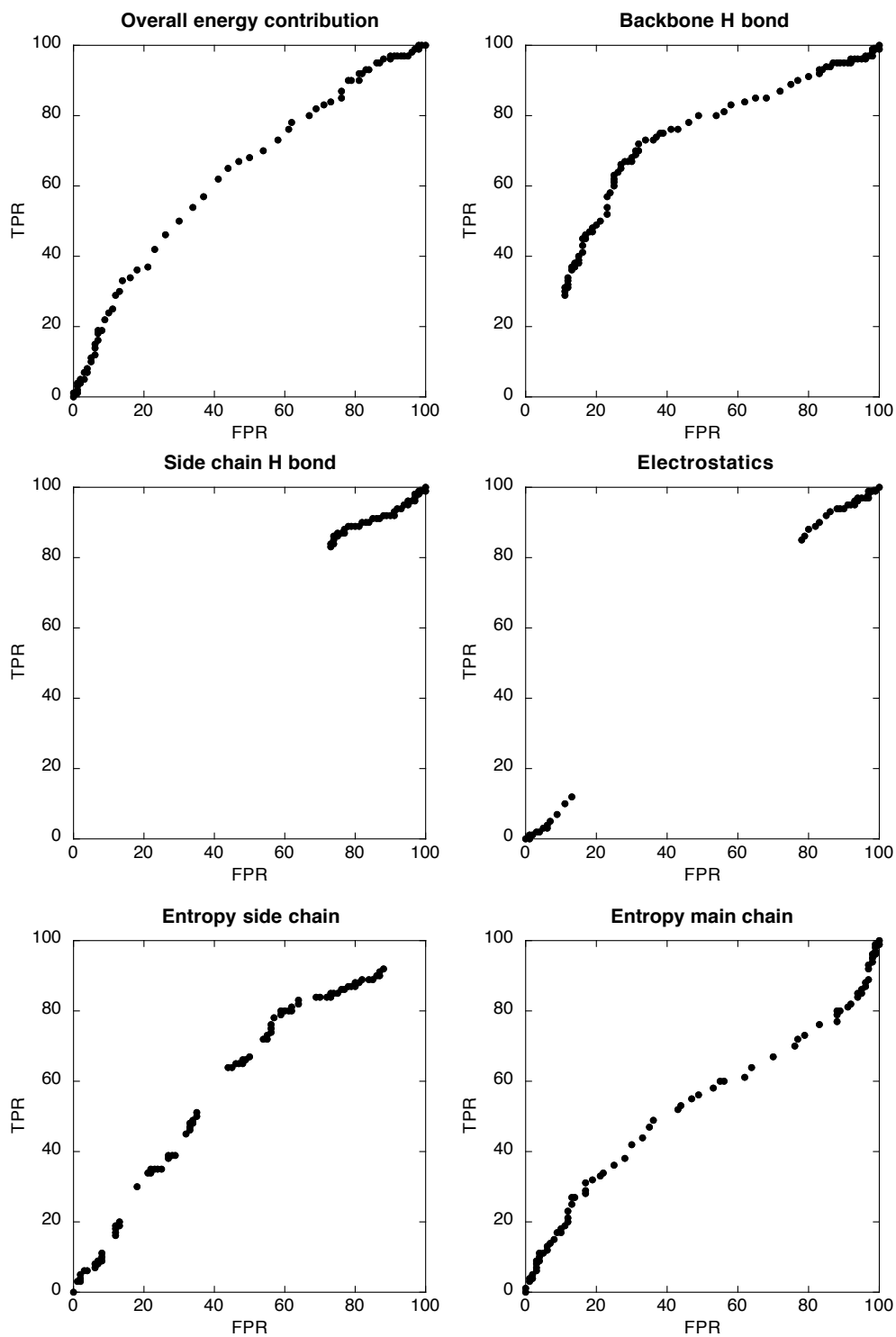
Figure S2. ROC curves for classification of disease mutations and neutral variation by using structural properties of the amino acid substitution site.
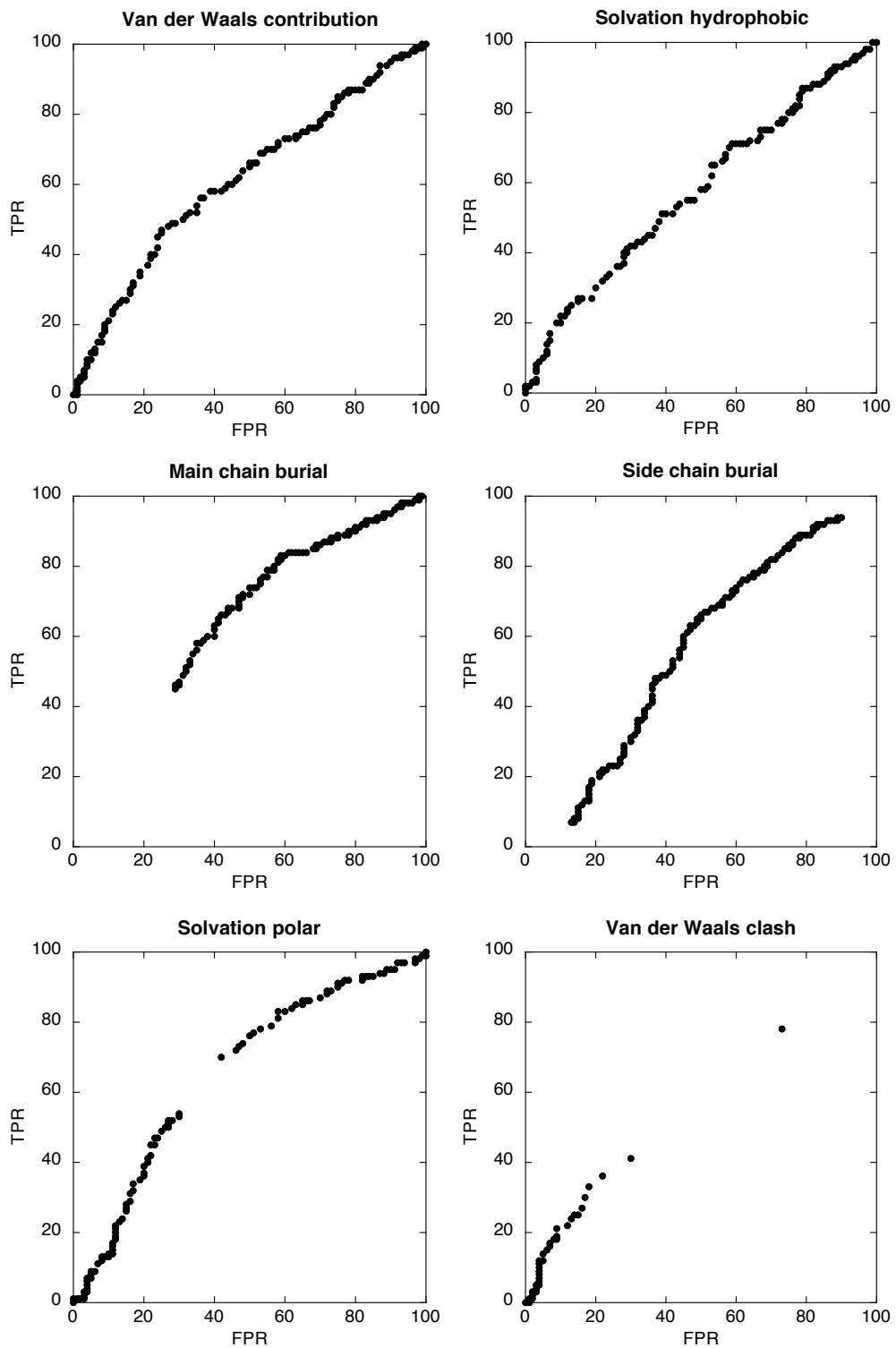
Figure S2 *(continued).* ROC curves for classification of disease mutations and neutral variation by using structural properties of the amino acid substitution site.
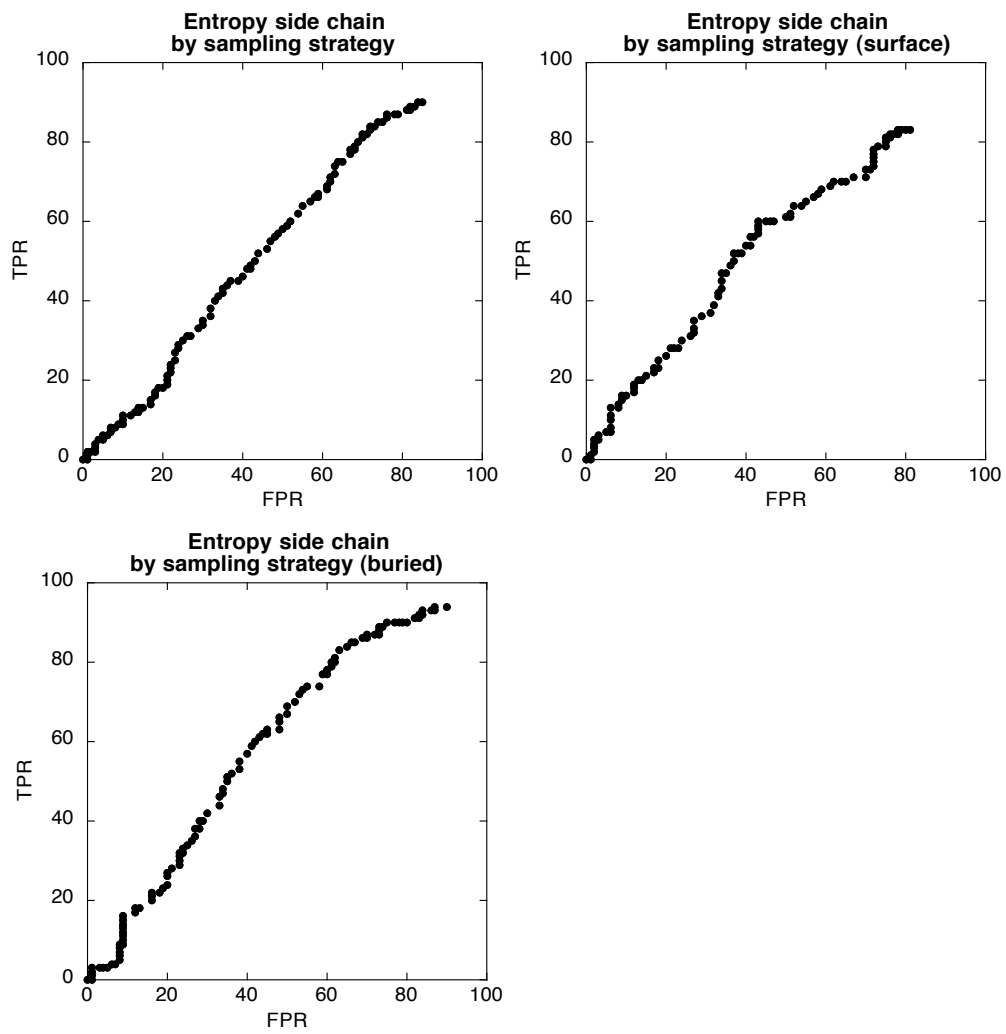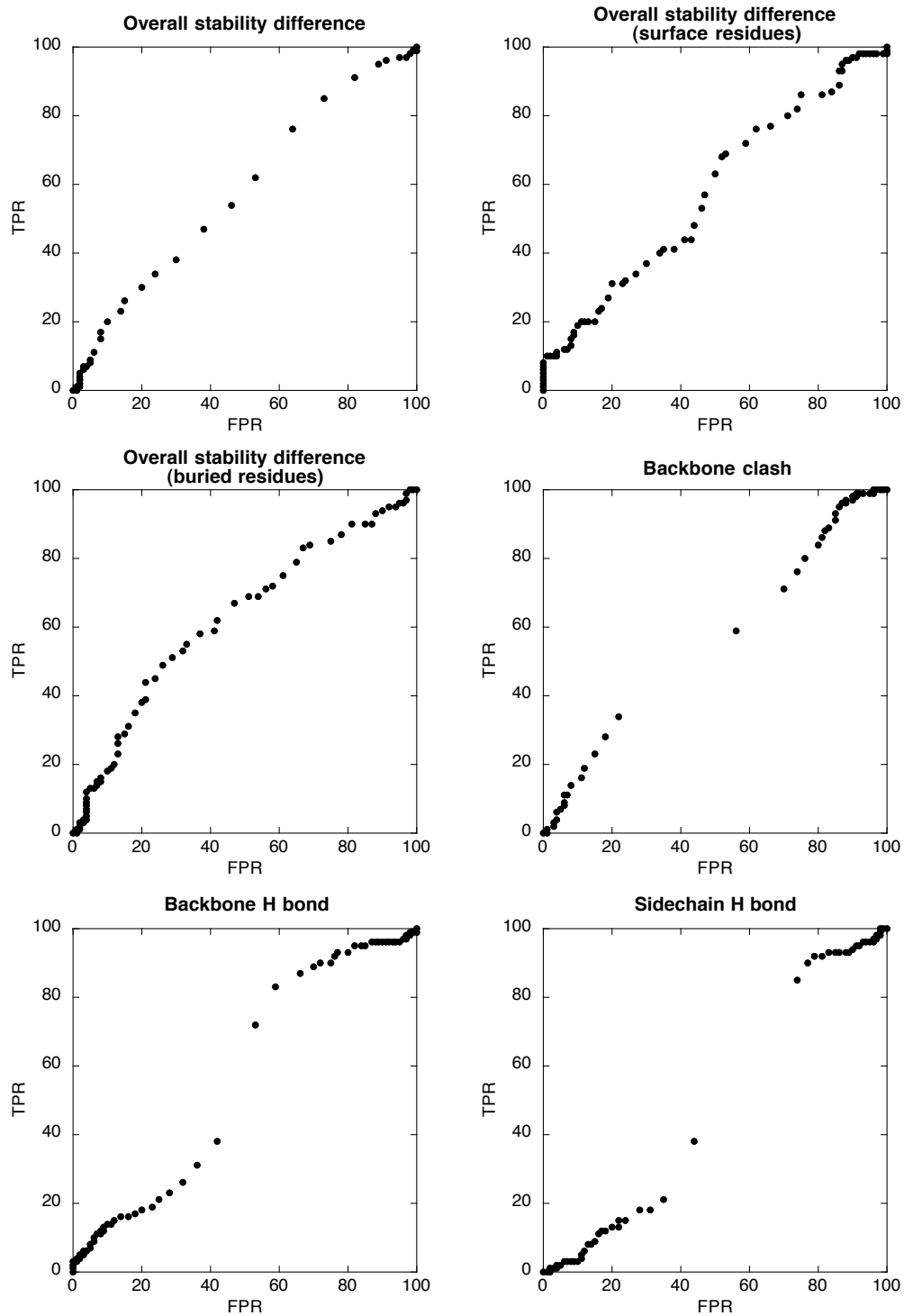
Figure S2 *(continued).* ROC curves for classification of disease mutations and neutral variation by using structural properties of the amino acid substitution site.

Figure S3. ROC curves for classification of disease mutations and neutral variation by using structural differences between the wild type and variant protein.
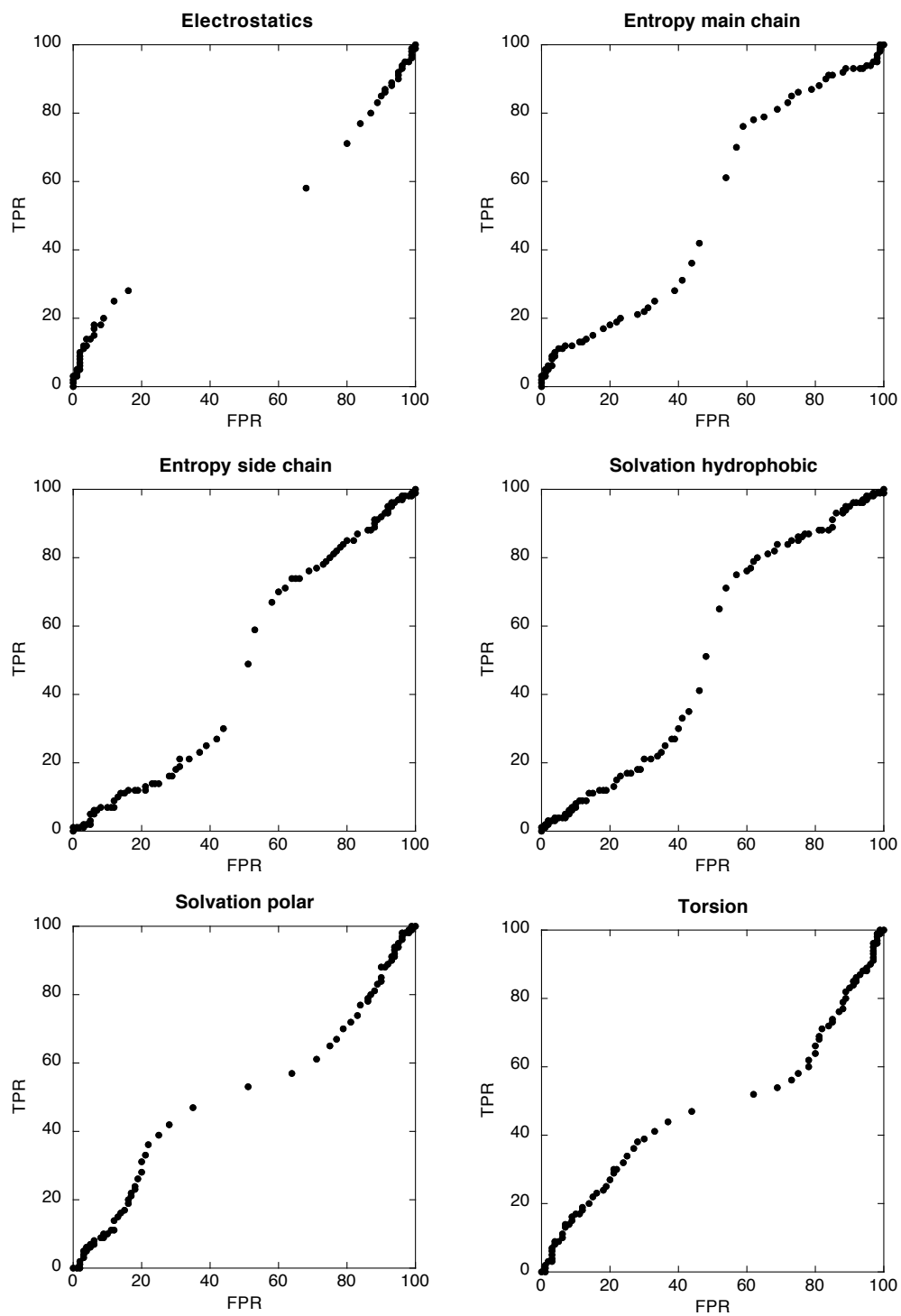
Figure S3 *(continued).* ROC curves for classification of disease mutations and neutral variation by using structural differences between the wild type and variant protein.
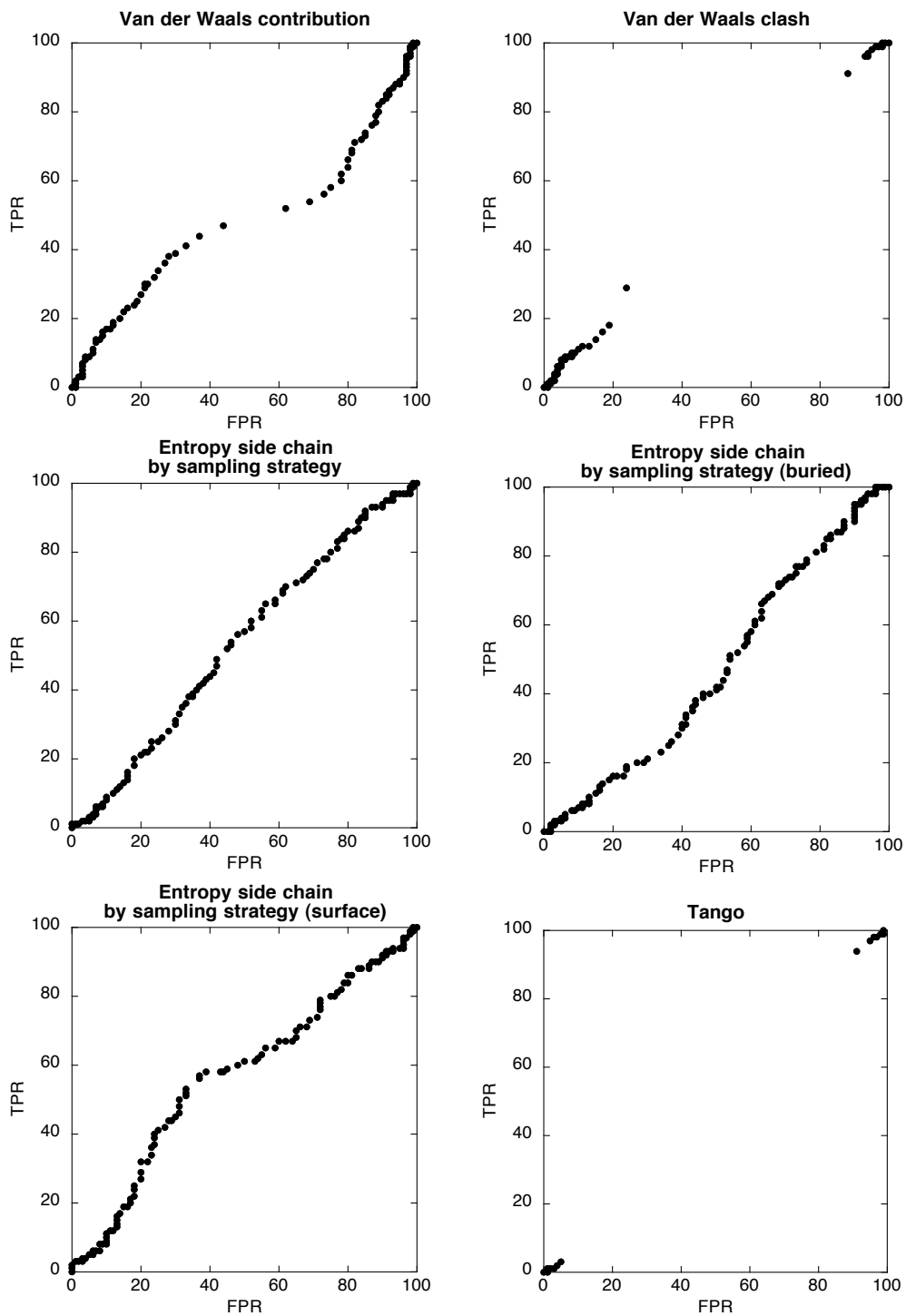
Figure S3 *(continued).* ROC curves for classification of disease mutations and neutral variation by using structural differences between the wild type and variant protein.
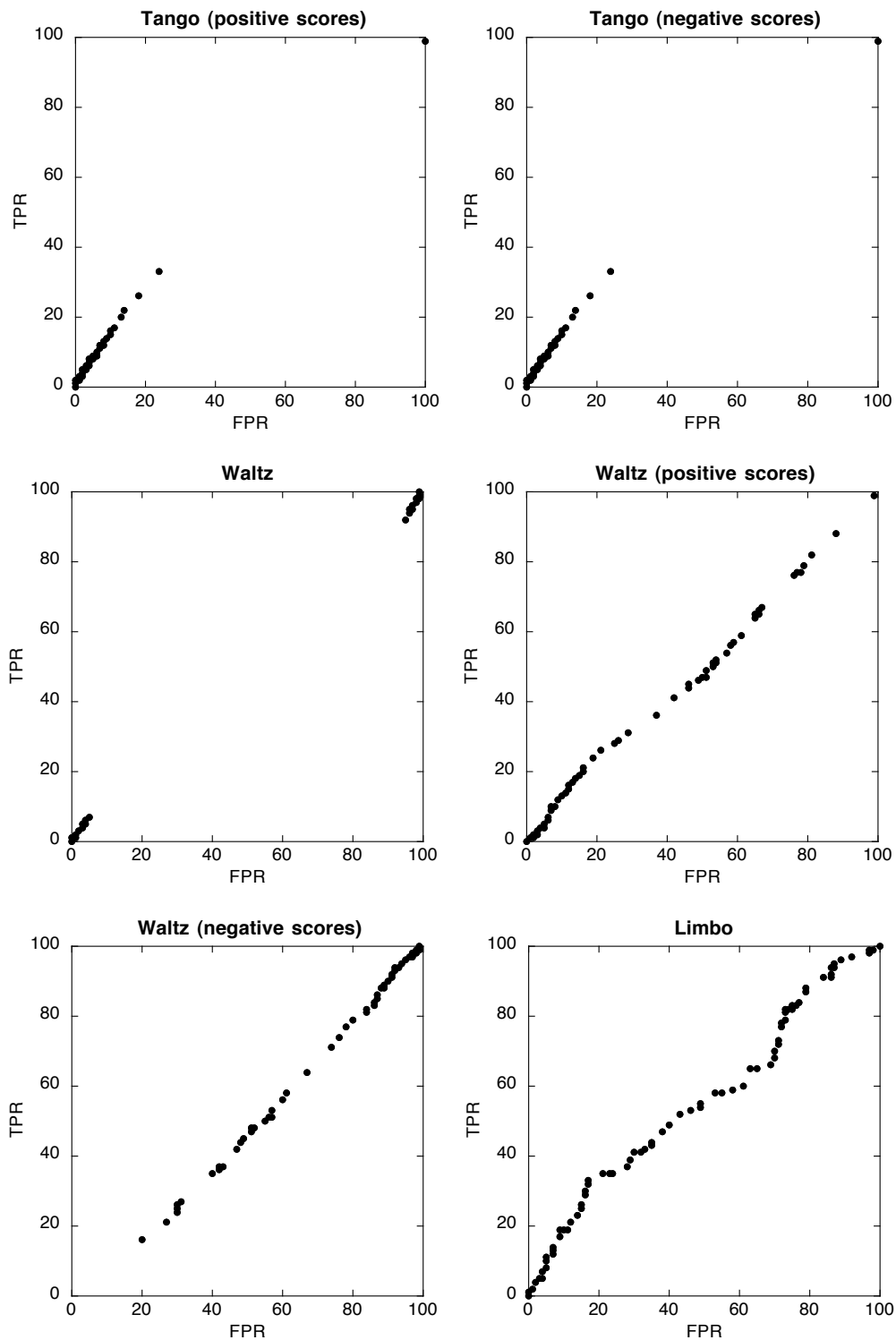
Figure S3 *(continued).* ROC curves for classification of disease mutations and neutral variation by using structural differences between the wild type and variant protein.

References

1. Chasman D, Adams RM: **Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: Structure-based assessment of amino acid variation**. *J Mol Biol* 2001, **307**(2):683–706.

2. Clifford RJ, Edmonson MN, Nguyen C, Buetow KH: **Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms**. *Bioinformatics* 2004, **20**(7):1006–1014.

3. Ferrer-Costa C, Orozco M, de la Cruz X: **Sequence-based prediction of pathological mutations**. *Proteins* 2004, **57**(4):811–819.

4. Jiang R, Yang H, Sun F, Chen T: **Searching for interpretable rules for disease mutations: a simulated annealing bump hunting strategy**. *BMC Bioinformatics* 2006, **7**:417.

5. Krishnan VG, Westhead DR: **A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function**. *Bioinformatics* 2003, **19**(17):2199–2209.

6. Needham CJ, Bradford JR, Bulpitt AJ, Care MA, Westhead DR: **Predicting the effect of missense mutations on protein function: analysis with Bayesian networks**. *BMC Bioinformatics* 2006, **7**:405.

7. Ng PC, Henikoff S: **Predicting deleterious amino acid substitutions**. *Genome Res* 2001, **11**(5):863–874.

8. Saunders CT, Baker D: **Evaluation of structural and evolutionary contributions to deleterious mutation prediction**. *J Mol Biol* 2002, **322**(4):891–901.

9. Yue P, Li Z, Moult J: **Loss of protein structure stability as a major causative factor in monogenic disease**. *J Mol Biol* 2005, **353**(2):459–473.

10. Ferrer-Costa C, Orozco M, de la Cruz X: **Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties**. *J Mol Biol* 2002, **315**(4):771–786.

11. Bao L, Cui Y: **Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information**. *Bioinformatics* 2005, **21**(10):2185–2190.

12. Bao L, Cui Y: **Functional impacts of non-synonymous single nucleotide polymorphisms: Selective constraint and structural environments**. *FEBS Lett* 2006, **580**(5):1231–4.

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

14. Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A: **LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources**. *Bioinformatics* 2005, **21**(12):2814–2820.

15. Stitziel NO, Tseng YY, Pervouchine D, Goddeau D, Kasif S, Liang J: **Structural location of disease-associated single-nucleotide polymorphisms**. *J Mol Biol* 2003, **327**(5):1021–1030.

16. Worth CL, Bickerton GRJ, Schreyer A, Forman JR, Cheng TMK, Lee S, Gong S, Burke DF, Blundell TL: **A structural bioinformatics approach to the analysis of nonsynonymous single nucleotide polymorphisms (nsSNPs) and their relation to disease**. *J Bioinform Comput Biol* 2007, **5**(6):1297–1318.

17. Burke DF, Worth CL, Priego EM, Cheng T, Smink LJ, Todd JA, Blundell TL: **Genome bioinformatic analysis of nonsynonymous SNPs**. *BMC Bioinformatics* 2007, **8**:301.

18. Worth CL, Burke DF, Blundell TL: **Estimating the effects of single nucleotide polymorphisms on protein structure: how good are we at identifying likely disease associated mutations?** In *Proceedings of Molecular Interactions - Bringing Chemistry to Life* 2006.

19. Ng PC, Henikoff S: **Accounting for human polymorphisms predicted to affect protein function**. *Genome Res* 2002, **12**(3):436–446.

20. Wang Z, Moult J: **SNPs, protein structure, and disease**. *Hum Mutat* 2001, **17**(4):263–270.

21. Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A: **Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis**. *Nat Genet* 1999, **22**(3):239–247.

22. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES: **Characterization of single-nucleotide polymorphisms in coding regions of human genes (vol 22, pg 231, 1999)**. *Nat Genet* 1999, **23**(3):373–373.

23. Serrano L, Guerois R: **Fold-X: An algorithm to predict and engineer folding pathways**. *Abstr Pap Am Chem Soc* 2001, **221**:U395–U395.

24. Guerois R, Nielsen JE, Serrano L: **Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations**. *J Mol Biol* 2002, **320**(2):369–387.

25. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L: **The FoldX web server: an online force field**. *Nucleic Acid Res* 2005, **33**:W382–W388.

26. Schymkowitz JWH, Rousseau F, Martins IC, Ferkinghoff-Borg J, Stricher F, Serrano L: **Prediction of water and metal binding sites and their affinities by using the Fold-X force field**. *Proc Natl Acad Sci USA* 2005, **102**(29):10147–10152.

27. Vriend G: **What If - a molecular modeling and drug design program**. *J Mol Graph* 1990, **8**:52–.

28. Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L: **Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins**. *Nat Biotechnol* 2004, **22**(10):1302–1306.

29. Maurer-Stroh S, Kuemmerer N, Lopez de la Paz M, Martins I, Reumers J, Serrano L, Rousseau F, Schymkowitz J: **Accurate prediction of sequence determinants of amyloid formation using the Waltz algorithm**. *Submitted* 2007.

30. Thompson MJ, Sievers SA, Karanicolas J, Ivanova MI, Baker D, Eisenberg D: **The 3D profile method for identifying fibril-forming segments of proteins.** *Proc Natl Acad Sci U S A* 2006, **103**(11):4074–4078.

31. Van Durme J, Maurer-Stroh S, Wilkinson H, Rousseau F, Schymkowitz J: **Accurate prediction of the sequence determinants of DnaK-peptide binding via a method that integrates homology modelling and experimental data**. *Submitted* 2007.

32. McCarty JS, Rudiger S, Schonfeld HJ, Schneider-Mergener J, Nakahigashi K, Yura T, Bukau B: **Regulatory region C of the E. coli heat shock transcription factor, sigma32, constitutes a DnaK binding site and is conserved among eubacteria**. *J Mol Biol* 1996, **256**(5):829–37.