SigniSite: Identification of residue-level genotype-phenotype correlations in protein multiple sequence alignments

Jessen, Leon Eyrich; Hoof, Ilka; Lund, Ole; Nielsen, Morten

Published in:
Nucleic Acids Research

Link to article, DOI:
10.1093/nar/gkt497

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

W286–W291 *Nucleic Acids Research, 2013, Vol. 41, Web Server issue* *Published online 12 June 2013*
doi:10.1093/nar/gkt497

SigniSite: Identification of residue-level genotype-phenotype correlations in protein multiple sequence alignments

Leon Eyrich Jessen[1], Ilka Hoof[2], Ole Lund[1] and Morten Nielsen[1,3,*]

[1]Department of Systems Biology, Center for Biological Sequence Analysis, Technical University of Denmark, Kemitorvet, Building 208, DK-2800 Lyngby, Denmark, [2]Department of Molecular Biology and Biotech Research and Innovation Centre (BRIC), Bioinformatics Centre, University of Copenhagen, Ole Maaloes Vej 5, DK-2200 Copenhagen, Denmark and [3]Instituto de Investigaciones Biotecnologicas, Universidad Nacional de San Martin, San Martin, B 1650 HMP, Buenos Aires, Argentina

Received January 31, 2013; Revised May 2, 2013; Accepted May 15, 2013

ABSTRACT

Identifying which mutation(s) within a given genotype is responsible for an observable phenotype is important in many aspects of molecular biology. Here, we present *SigniSite*, an online application for subgroup-free residue-level genotype–phenotype correlation. In contrast to similar methods, *SigniSite* does not require any pre-definition of subgroups or binary classification. Input is a set of protein sequences where each sequence has an associated real number, quantifying a given phenotype. *SigniSite* will then identify which amino acid residues are significantly associated with the data set phenotype. As output, *SigniSite* displays a sequence logo, depicting the strength of the phenotype association of each residue and a heat-map identifying 'hot' or 'cold' regions. *SigniSite* was benchmarked against SPEER, a state-of-the-art method for the prediction of specificity determining positions (SDP) using a set of human immunodeficiency virus protease-inhibitor genotype–phenotype data and corresponding resistance mutation scores from the Stanford University HIV Drug Resistance Database, and a data set of protein families with experimentally annotated SDPs. For both data sets, *SigniSite* was found to outperform SPEER. *SigniSite* is available at: http://www.cbs.dtu.dk/services/SigniSite/.

INTRODUCTION

Whether conducting research in vaccine design or trying to elucidate the intimate details of a given receptor::ligand interaction, genotype–phenotype correlation is a powerful tool to enhance the understanding of the minute subtleties, often characterizing research within the field of molecular biology.

The traditional approach for wet-laboratory analysis of genotype–phenotype correlations involves site-directed mutagenesis and subsequent quantification of mutation-impact on the phenotype, e.g. binding-affinity or catalytic efficiency. This approach of mutating all amino acid residues in a given protein is a time consuming and tedious task. Random mutagenesis has the advantage of introducing a large number of random mutations throughout the protein. One example of application of random mutagenesis is to increase the signal from near-infrared fluorescent proteins (1). In such a panel of sequenced variants with multiple mutations, it is a complex task to systematically pinpoint the exact amino acid residue(s), i.e. the genotype, associated with a given phenotype (e.g. fluorescence). Another area of application is genotype–phenotype association studies in proteins, which show inherent natural variability, as is the case for instance for proteins involved in the pathogenesis of malaria (2).

Here, we present *SigniSite*, an online application for subgroup-free residue-level genotype–phenotype correlation in protein multiple sequence alignments (MSAs). A number of methods have been developed for the identification of functional sites in protein sequences (3–10), most requiring a definition of functional subgroups before analysis. However, if the phenotype associated with the sequences is not categorical (e.g. substrate-specificity) but continuous (e.g. catalytic efficiency), a pre-division of sequences subgroups is none trivial. In contrast, *SigniSite* does not require any subgroup division or binary classification. Instead, *SigniSite* directly analyses the raw sequences and associated continuous values. The main novelty of *SigniSite* is that unlike conventional methods for the prediction of specificity determining

*To whom correspondence should be addressed. Tel: +45 45 25 61 27; Fax: +45 45 93 15 85; Email: jessen@cbs.dtu.dk

positions (SDP), it not only predicts the positions in the MSA determining a given protein function but also makes a statistical evaluation of which types of amino acid residue substitutions (genotype) are associated with the observable phenotype at the SDP.

The web server implementation of the *SigniSite* method described here is an automatized online application with an easy-to-interpret graphical output. The application is easy to use for the non-expert end-user and aims at aiding researchers in the analysis of sequence data, where the phenotype is quantified by a real number. A list of abbreviations is available in the Supplementary Data.

THE WEB SERVER

User interface

The *SigniSite* server is intended to provide the non-expert user with a simple interface. At default settings, an amino acid residue is considered significantly associated with the MSA phenotype, if the *P*-value for the specific residue is smaller than or equal to $\alpha = 0.05$ after Bonferroni Single-Step Correction for Multiple Testing (CMT) (11). On the submission page, sequences can be submitted to the server either as paste-in or via the file upload field. On submission, *SigniSite* will check whether the submitted sequences are aligned. If not, an MSA will be created using MAFFT (12). *SigniSite* will exclude any characters other than the one-letter representation of the 20 standard proteogenic amino acids from the analysis.

Input

As input *SigniSite* takes an MSA in FASTA-format (minimum two sequences). Each sequence must have an associated real number, stated white-space-separated as the last element in its FASTA header. At least two different values must exist in the MSA. The MSA is assumed pre-sorted, if the end-placed value is absent. A section with options for customizing the analysis is available. The following parameters are user-adjustable: (i) the level of significance 'α', $0 \leq \alpha \leq 1$ (default is 0.05). (ii) The method for CMT: 'Bonferroni Single-Step' (default), 'Holm Step-Down' (11) or 'no correction'. (iii) The sorting of the sequences: 'Decreasing', highest sequence-associated value is considered the strongest, e.g. fluorescent protein signals, and vice versa for 'Increasing', e.g. binding affinity. Furthermore, the user can choose a reference sequence to assign sequence-specific positional output numbering. This is useful, when the MSA contains insertions. Finally, the user can modify the logo output by choosing to include either 'Significant positions' (default, displays all residues at positions where at least one amino acid residue has been identified as significantly associated with the data set phenotype), 'Significant Residues' (as for significant positions, but only including significant residues) or 'Full Logo' (all residues at all positions). At the results page, a button below the generated logo allows the user to fully customize the logo using Seq2Logo (13).

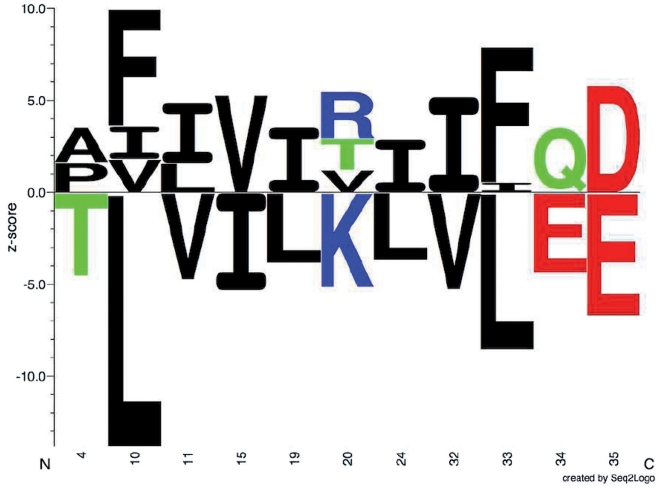

Figure 1. Sequence logo. Example of sequence logo (13) output from *SigniSite* from the analysis of the ATV ~Antivirogram multiple sequence alignment (MSA), truncated to p_1–p_{35} for the purpose of illustration (see 'Materials and Methods' section). The analysis was performed with default settings. On the *x*-axis are the MSA positions p and on the *y*-axis the *Z*-scores for each amino acid residue a ($z_{p,a}$). The height of each letter representing the residues is proportional to $z_{p,a}$, i.e. the strength of the statistical association between the residue and the data set-phenotype. Residues above the $Z = 0$ line have a $z_{p,a} > 0$, i.e. enhances the phenotype, whereas residues below the $Z = 0$ line have a $z_{p,a} < 0$, i.e. inhibits the phenotype, e.g. the presence of a certain residue with favourable chemical properties may enhance binding ($z_{p,a} > 0$), whereas a residue with unfavourable properties may inhibit binding ($z_{p,a} < 0$). Colour-coding: acidic [DE]: red, basic [HKR]: blue, hydrophobic [ACFILMPVW]: black and neutral [GNQSTY]: green (14).

Output

The *SigniSite* output is intended to provide the end-user with an easily interpretable graphical representation of the statistical evaluations performed by *SigniSite*. An example of a sequence logo (13) generated by *SigniSite* is shown in Figure 1. The logo gives an overview of residue associations. See Figure 1 legend for further details. *SigniSite* will also generate a heatmap (Figure 2). The heatmap is intended to give a graphic overview of 'hot' and 'cold' regions in the MSA, with respect to the data set phenotype. See Figure 2 legend for details.

RESULTS

As an initial performance evaluation, we chose to analyse 18 human immunodeficiency virus type 1 (HIV-1) MSAs compiled from the Stanford University HIV Drug Resistance Database (15,16) (HIVdb) using Spearman's rank correlation (SCC) to correlate the obtained *SigniSite* *Z*-scores ($z_{p,a}$ for each residue a at each position p) with the table of resistance mutation scores (RMS) also available from the HIVdb (see 'Materials and Methods' section), i.e. $\mathrm{SCC}(z_{p,a} \sim \mathrm{RMS})$. Results are given in Table 1.

As the SCC evaluation is threshold dependent, a threshold-independent performance evaluation was added using the area under the receiver operator

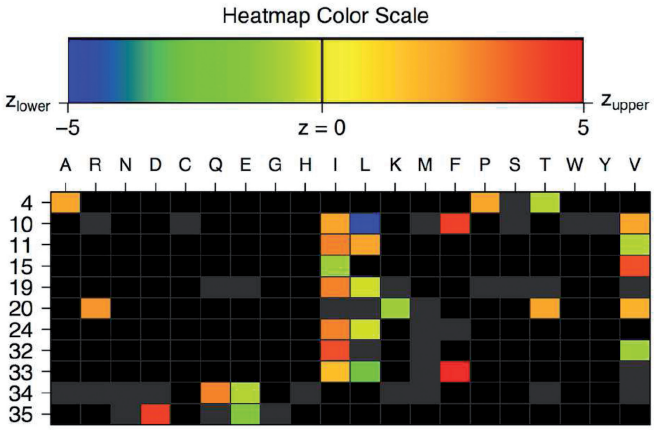

Figure 2. *SigniSite* heatmap from the analysis of the ATV ~Antivirogram multiple sequence alignment (MSA), truncated to $p_1 - p_{35}$ for the purpose of illustration (see 'Materials and Methods' section). The analysis was performed with default settings. On the *x*-axis are the 20 proteogenic amino acids *a* and on the *y*-axis the positions *p* in the analysed MSA. The colour coding of the fields is such that fields reflecting $z_{p,a} \leq -5$ are blue, whereas $z_{p,a} \geq 5$ results in a red field. For $-5 < z_{p,a} < 5$, nuances in between are used. If a residue has a $z_{p,a}$ of 0, the cell is coloured grey. Absent residues are coloured black. If only one grey cell is present at a given position, this implies that the position is fully conserved, harbouring only this residue. If more grey cells are present, their associated *P*-values have become $P = 1 \Rightarrow z_{p,a} = 0$ after correction for multiple testing.

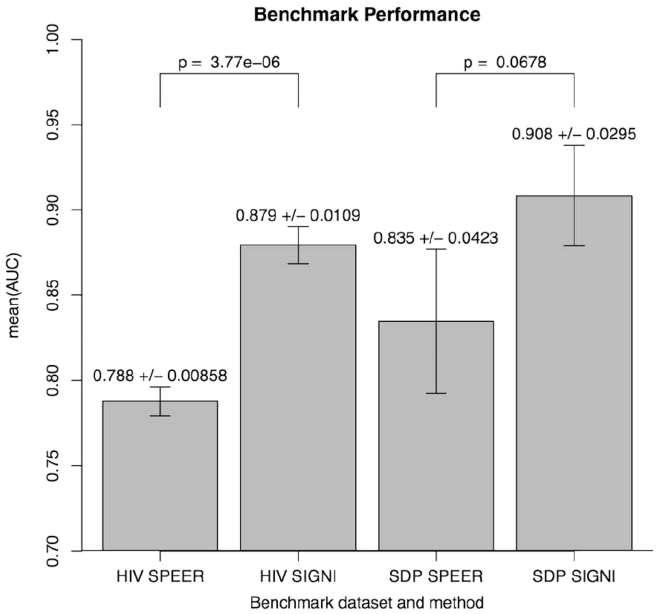

Figure 3. Measures are mean (AUC) ± SE. Columns are: HIV [SPEER/SIGNI], SPEER and *SigniSite's* predictions on the HIVdb data set. SDP [SPEER/SIGNI] SPEER and *SigniSite's* predictions on the SDP data set. *P*-values quantifying the significance of the difference in performance were obtained using a two-tailed paired *t*-test.

Table 1. Benchmark results

Measure	$\|z\| \geq 0$	$\|z\| \geq 1.96$	$\|z\| \geq 1.96_{CMT}$
SCC[a]	0.451 ± 0.015	0.506 ± 0.016	0.542 ± 0.020
MCC[b]	0.492 ± 0.028	0.387 ± 0.027	0.297 ± 0.040
SENS[b]	0.915 ± 0.015	0.598 ± 0.056	0.386 ± 0.055
SPEC[b]	0.579 ± 0.016	0.774 ± 0.031	0.882 ± 0.022

[a]Calculated against the RMS.
[b]Calculated against the $(RMS + IAS)_{mut}$.
Measures are means ± SE. CMT: corrected for multiple testing, SCC: Spearman's rank correlation, MCC: Matthews Correlation Coefficient, SENS: sensitivity, SPEC: specificity.

characteristics curve (AUC) measure, resulting in $AUC(z_{p,a} \sim RMS_{bin}) = 0.791 \pm 0.010$. Certain mutations not included in the RMS were repeatedly identified by *SigniSite*. As the majority of these mutations were found in the binary resistance annotations from the international antiviral society-USA (IAS) (17), we enriched the RMS_{bin} with the IAS and re-calculated the AUC, obtaining a significant performance increase of $AUC(z_{p,a} \sim (RMS{+}IAS)_{mut}) = 0.822 \pm 0.011 (P = 5.16 \cdot 10^{-4})$, two-tailed paired *t*-test).

Furthermore, we evaluated the performance of *SigniSite* using performance measures: Matthew's correlation coefficient (MCC), sensitivity (SENS) and specificity (SPEC) against $(RMS + IAS)_{mut}$. See Table 1 for results.

Having obtained good results for both the threshold-dependent and -independent performance evaluations, we turned to benchmark *SigniSite* against similar existing methods. In a 2009 benchmark study (18), SPEER (5,19) was identified as the state-of-the-art method for prediction of specificity definition positions (SDP). We, therefore, here compared the performances of *SigniSite* and SPEER on each of their original benchmarks data sets (see 'Materials and Methods' section) against $(RMS + IAS)_{pos}$. The results are shown in Figure 3. The results show that *SigniSite* outperforms SPEER on both data sets. The difference in predictive performance was, however, only found to be statistically significant for the HIVdb data set.

DISCUSSION

SigniSite aims at providing a simple-to-use method for subgroup-free residue-level genotype–phenotype correlation in protein MSAs. *SigniSite*, thus, addresses a long-existing challenge in molecular biology; genotype-phenotype mapping. Genotype–phenotype mapping has a wide range of purposes in molecular biology, e.g. structural regions responsible for immunity (2), identifying protein-variants responsible for the severity of a disease (20) or coupling receptor polymorphisms to surface expression (21) etc.

Site-directed mutagenesis in proteins and subsequent quantification of mutation-impact on a given phenotype is a time consuming and tedious task. High-throughput methods such as e.g. random mutagenesis (1) have, therefore, been developed. However, the challenge of analysing the increasingly larger volumes of data being generated only becomes greater. Additionally, large genotype–phenotype data sets (GPDs) can be compiled from publicly available databases, such as the HIVdb (15,16). *SigniSite* addresses this exact challenge.

SigniSite was benchmarked on publicly available GPDs and RMS from the Stanford University HIV Drug Resistance Database (HIVdb) (15,16). We observed that for each of the 18 different benchmark data sets, *SigniSite* consistently identified certain residues, not annotated in the RMS table, as significantly associated with anti-viral drug resistance. We compared these identifications with binary resistance annotations from the International Antiviral Society-USA (IAS) (17) and found that the majority were indeed annotated as resistance impacting. This observation suggests that the RMS data are not exhaustive, and that the obtained correlation should rather be regarded as a lower bound of the true predictive performance.

As the SDP method SPEER (5,19) was found to be the state-of-the-art method in a 2009 benchmark study (18), we chose to compare *SigniSite* to SPEER. We observed that *SigniSite* significantly outperformed SPEER on the HIVdb data set ($P = 3.77 \cdot 10^{-6}$) and for the SDP data set (as defined in the SPEER paper), *SigniSite* likewise outperformed SPEER, approaching a significant difference ($P = 0.0678$). Furthermore, *SigniSite* was much faster, taking only a few minutes to analyse the largest of the MSA ($n_{seqs} = 1{,}374$). SPEER on the other hand requires to be compiled in a slower version, when $n_{seqs} > 200$, taking $\sim$2 h to complete the analysis.

In conclusion, *SigniSite* provides two important novel features: (i) *SigniSite* does not require any manual annotation of the data before analysis, e.g. binder/non-binder classification, *SigniSite* requires only sequences and associated values. (ii) Unlike conventional SDP prediction methods like SPEER, *SigniSite* will not only identify positions impacting the phenotype but also pinpoint the exact amino acid residue substitution(s) responsible for the impact detected at the identified position. To the best of our knowledge, this level of resolution has so far not been available.

MATERIALS AND METHODS

Benchmark data sets

Summary, see Supplementary Data for details.

HIVdb resistance mutation scores

The table of RMS was downloaded from the HIVdb (15,16), available at http://hivdb.stanford.edu/DR/cgi-bin/rules_scores_hivdb.cgi?class=PI. The table of RMS contains information about positions known to harbour mutations ($n = 688$) compared with wild-type (WT) and their impact on resistance towards eight different protease inhibitors (PIs). Positive scores range is [3,60] ($n = 296$) and indicates that the mutation increases the resistance towards a given PI. Negative score range is [−5, −10] ($n = 15$) and indicates a decreased resistance. Scores of 0 ($n = 377$) indicate lack of resistance impact. At each position annotated in the table of RMS, the consensus residue was assigned an RMS of 0.

IAS resistance annotations

Protease mutations known to impact PI resistance were retrieved from the table 'mutations in the protease gene associated with resistance to protease inhibitors', in the International Antiviral Society USA (IAS)'s Update of the Drug Resistance Mutations in HIV-1: March 2013 (17). Also here, the consensus residue at annotated resistance positions was assigned an IAS score of 0.

Table transformations

The following table transformations were performed: $\text{RMS} \rightarrow \text{RMS}_{bin}$, such that $\text{RMS} > 0 \Rightarrow \text{RMS}_{bin} = 1$, otherwise $\text{RMS}_{bin} = 0$. RMS_{bin}+IAS $\rightarrow (\text{RMS+IAS})_{mut}$, such that $\text{RMS}_{bin} > 0$ or $\text{IAS} > 0 \Rightarrow (\text{RMS+IAS})_{mut} = 1$, otherwise $(\text{RMS+IAS})_{mut} = 0$. $(\text{RMS+ IAS})_{pos}$, such that for each position in $(\text{RMS} + \text{IAS})_{mut}$ the resulting $(\text{RMS+IAS})_{pos} = 1$ if at least one $(\text{RMS+IAS})_{mut} > 0$, otherwise $(\text{RMS+IAS})_{pos} = 0$. In all tables, any score $s_{table} > 0$ is considered an actual positive and any score $s_{table} \leq 0$ is considered an actual negative (Table 2).

MSAs from the HIVdb protease GPDs

GPDs were downloaded from the Stanford University HIV Drug Resistance Database (HIVdb) (15,16) Version 5.0, March, 2012, available at http://HIVdb.stanford.edu/cgi-bin/GenoPhenoDS.cgi. MSAs were compiled from the GPDs. Each MSA contains the sequences of a set of HIV-1 protease variants with measured fold change in resistance (compared with WT) towards the same PI, measured using the same assay. Only PIs present in both the table of RMS and the GPDs were used limiting the analysis to 6 PIs: *ATV, IDV, LPV, NFV, SQV and TPV* each of which was assayed using the three assays: 'Antivirogram' (Virco™), 'PhenoSense' (ViroLogic™) and 'All Others'. A total of 12 714 sequences were constructed and compiled into 18 MSAs. The length of each of the protease variants is 99 amino acid residues.

The SPEER program and SDP benchmark data

SPEER, MSAs and corresponding experimentally annotated specificity determining sites were downloaded from the SPEER repository available at: ftp://ftp.ncbi.nih.gov/pub/SPEER/ (5,19). We downloaded the latest curated version of the data as described by Chakrabarti and Panchenko (18).

The *SigniSite* method

The method takes a set of (protein) sequences as input. If the sequences are not aligned, *Signisite* will use MAFFT (12) to make an MSA from the input sequences. Subsequently, the sequences are ranked with respect to a real number associated with each sequence, e.g. the replicative capacity or catalytic efficiency. For each amino acid at each position in the MSA, a non-parametric test is performed to test whether the observed ranks deviate significantly from the expected ranks. CMT of the resulting *P*-values may be performed using Bonferroni single-step or Holm step-down procedures. The resulting *Z*-scores per residue are visualized in a logo plot and a heatmap.

Table 2. Overview of target table notation

Notation	Format	Level	Annotating
RMS[a]	Real num.	Residue	Fold-change in PI resistance
IAS[b]	Binary	Residue	PI ass. resistance mutations
RMS_{bin}[c]	Binary	Residue	PI ass. resistance mutations
$(RMS+IAS)_{mut}$[d]	Binary	Residue	PI ass. resistance mutations
$(RMS+IAS)_{pos}$[e]	Binary	Position	Positions ass. with PI resistance

[a]It is used when calculating SCC, [b]it is used to look up mutations not annotated in **1**, but repeatedly identified by *SigniSite*, [c]it is used when calculating AUC, [d]it is used for the enriched AUC calculation and when calculating the MCC, SENS and SPEC, [e]it is used as positional targets, when comparing the predictive performances of *SigniSite* and SPEER.
'num.', 'ass.', 'PI' abbreviates 'numbers', 'association' and 'protease inhibitor'. In all tables, any score $s_{table} > 0$ is considered an actual positive and any score $s_{table} \leq 0$ is considered an actual negative.

Brief description of the method underlying SigniSite

(see Supplementary Data for details). Initially each sequence is assigned a rank by sorting the sequence associated values (either ascending or descending depending on type of value) and then assigning a rank of '1' to the first sequence after sorting, '2' to the second and so forth. Each amino acid residue a observed at position p ($res_{p,a}$) in the MSA is then assigned the rank of the sequence to which it belongs. This way each $res_{p,a}$ is associated with a specific rank. At each position in the MSA, the mean ranks of each residue type are then calculated and placed in a rank matrix, where each row corresponds to a position in the MSA and each column to one of the 20 standard proteogenic amino acids, sorted according to A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y and V (*SigniSite* will exclude any characters but these 20).

Subsequently, *SigniSite* evaluates for each position and residue type the difference between the mean of the observed and expected ranks. The mean of the expected ranks is the mean of the ranks we would observe if the residue type $res_{p,a}$ was randomly distributed over the column p in the MSA. This difference between observed and expected ranks is quantified by a Z-score assigned to each residue type at each position, yielding a Z-score-matrix. If a given position is fully conserved, $z = 0$ is assigned to the conserved residue. If a given residue type is absent at a given position, $z = 'NA'$ is assigned.

The non-parametric statistics, on which *SigniSite* is based, are similar to that of Wilcoxon test statistics (22), where the obtained evaluation scores can be approximated by the standard normal distribution, thus allowing Z-score conversion to P-values by standard method. As one test is performed per residue type, per position, *SigniSite* will by default apply Bonferroni single-step (11) CMT to adjust the reported P-values.

Benchmarking

For each of the 18 MSAs compiled from the HIVdb GPDs (see 'Materials and Methods' section), a set of predictions were made (Z-scores) estimating the strength of the association of each residue type a at each position p ($z_{p,a}$) to the phenotype of the MSA. The obtained set of $z_{p,a}$'s was then correlated with the RMS using Spearman's rank correlation (SCC) at three significance thresholds: including residues for which: (i) $P \leq 1$, (ii) $P \leq 0.05$ and (iii) $P \leq 0.05$ after CMT. The SCC was recorded for each of the 18 MSAs, and the mean and standard error (SE) of the means were calculated.

For evaluating threshold-independent performance, the AUC measure was applied. The AUC was calculated against two sets of targets: RMS_{bin} and the enriched set of targets $(RMS+IAS)_{mut}$. The mean AUC and SE were calculated for each set of targets.

Finally, the sensitivity, specificity and MCC were calculated at the same thresholds as the SCC against the enriched set of targets $(RMS+IAS)_{mut}$. The sensitivity, specificity and MCC were recorded for each of the 18 MSAs, and the means and SEs were calculated.

Comparing SigniSite and SPEER

To compare the performance of *SigniSite* with that of existing methods, we turned to a 2009 benchmark study by Chakrabarti and Panchenko (18) comparing the predictive performance of five SDP prediction methods, on a set of protein families with experimentally annotated SDPs. As SPEER (5,19) in this benchmark was found to be the best performing method, we here limit our analysis to comparing *SigniSite* and SPEER by applying both methods to their respective GPDs.

SPEER outputs positional predictions, whereas *SigniSite* assigns a Z-score for each residue type at each position. To cast the *SigniSite* Z-scores into one score per positions, the maximum of the absolute Z-scores was chosen.

SigniSite assigns a prediction value to all positions regardless of residue composition, whereas SPEER by default will skip any fully conserved and positions with >20% gaps. To get prediction values for all positions, we assign a value of '−100' to positions not predicted by SPEER (this value is lower than any score predicted by SPEER).

SPEER requires each sequence in an MSA to be subgroup-annotated before analysis. To accommodate this requirement, each HIV MSA was split into two subgroups, by sorting the sequences in the MSA descending on their associated real values and then splitting the sequences into subgroup '1' or '2' on the median of the sorted values.

To perform the rank analysis *SigniSite* requires that each sequence in the MSA has an associated real number. Of the 20 SDP MSAs, 13 contain only subgroups '1' and '2'. We chose to use these 13 MSAs for the benchmark, using '1' or '2' as '*SigniSite* real number values'.

This way the following two comparisons were made: *SigniSite* versus SPEER on the HIV protease data set and *SigniSite* versus SPEER in the SDP data set. The AUC measure was used to quantify the performance of each method on each benchmark data set.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary description of the *SigniSite* Method,

Supplementary descriptions of the benchmark data sets, Supplementary section on the impact of chosen seed for random number generation, Supplementary description of the benchmarks strategy, Supplementary Tables of HIV-1 PIs and abbreviations.

ACKNOWLEDGEMENTS

The authors thank Martin Blythe for coming up with the name *SigniSite*.

FUNDING

National Institutes of Health [HHSN272201200010C]; EU FP7 PepChipOmics: The European Union 7th Framework Program FP7/2007-2013 [222773]; The Center for Genomic Epidemiology (www.genomicepidemiology.org) grant 09-067103/DSF from the Danish Council for Strategic Research; The University of Copenhagen - Program of Excellence. Funding for open access charge: Technical University of Denmark - PhD programme.

Conflict of interest statement. None declared.

REFERENCES

1. Shcherbo,D., Shemiakina,I.I., Ryabova,A.V., Luker,K.E., Schmidt,B.T., Souslova,E.A., Gorodnicheva,T.V., Strukova,L., Shidlovskiy,K.M., Britanova,O.V. *et al*. (2010) Near-infrared fluorescent proteins. *Nat. Methods*, **7**, 827–829.
2. Gnidehou,S., Jessen,L., Gangnard,S., Ermont,C., Triqui,C., Quiviger,M., Guitard,J., Lund,O., Deloron,P. and Ndam,N.T. (2010) Insight into antigenic diversity of VAR2CSA-DBL5ε Domain from multiple *Plasmodium falciparum* placental isolates. *PLoS One*, **5**, e13105.
3. Brandt,B.W., Feenstra,K.A. and Heringa,J. (2010) Multi-Harmony: detecting functional specificity from sequence alignment. *Nucleic Acids Res.*, **38**, 35–40.
4. Capra,J.A. and Singh,M. (2008) Characterization and prediction of residues determining protein functional specificity. *Bioinformatics*, **24**, 1473–1480.
5. Chakrabarti,S., Bryant,S.H. and Panchenko,A.R. (2007) Functional specificity lies within the properties and evolutionary changes of amino acids. *J. Mol. Biol.*, **373**, 801–810.
6. Kalinina,O.V., Novichkov,P.S., Mironov,A.A., Gelfand,M.S. and Rakhmaninova,A.B. (2004) SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleic Acids Res.*, **32**, W424–W428.
7. Pei,J., Cai,W., Kinch,L.N. and Grishin,N.V. (2006) Prediction of functional specificity determinants from protein sequences using log-likelihood ratios. *Bioinformatics*, **22**, 164–171.
8. Ye,K., Feenstra,K.A., Heringa,J., Ijzerman,A.P. and Marchiori,E. (2008) Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a Machine-Learning approach for feature weighting. *Bioinformatics*, **24**, 18–25.
9. Buslje,C.M., Teppa,E., Domnico,T.D., Delfino,J.M. and Nielsen,M. (2010) Networks of high mutual information define the structural proximity of catalytic sites: implications for catalytic residue identification. *PLoS Comput. Biol.*, **6**, e1000978.
10. Lichtarge,O., Bourne,H.R. and Cohen,F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.
11. Dudoit,S., Yang,Y.H., Callow,M.J. and Speed,T.P. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sin.*, **12**, 111–139.
12. Katoh,K., Misawa,K., Kuma,K.I. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
13. Thomsen,M.C.F. and Nielsen,M. (2012) Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res.*, **40**, W281–W287.
14. Lund,O., Nielsen,M., Lundegaard,C., Kesmir,C. and Brunak,S. (2005) *Immunological Bioinformatics*. The MIT Press, Cambridge, MA, London, England.
15. Rhee,S.Y., Gonzales,M.J., Kantor,R., Betts,B.J., Ravela,J. and Shafer,R.W. (2003) Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.*, **30**, 298–303.
16. Shafer,R.W. (2006) Rationale and uses of a public HIV drug-resistance database. *J. Infect. Dis.*, **194**, S51–S58.
17. Johnson,V.A., Calvez,V., Gnthard,H.F., Paredes,R., Pillay,D., Shafer,R., Wensing,A.M. and Richman,D.D. (2013) Update of the drug resistance mutations in HIV-1: March 2013. *Top Antivir. Med.*, **21**, 6–14.
18. Chakrabarti,S. and Panchenko,A.R. (2009) Ensemble approach to predict specificity determinants: benchmarking and validation. *BMC Bioinformatics*, **373**, 801–810.
19. Chakraborty,A., Mandloi,S., Lanczycki,C.J., Panchenko,A.R. and Chakrabarti,S. (2012) SPEER-SERVER: a web server for prediction of protein specificity determining sites. *Nucleic Acids Res.*, **40**, W242–W248.
20. Healy,D.G., Falchi,M., O'Sullivan,S.S., Bonifati,V., Durr,A., Bressman,S., Brice,A., Aasly,J., Zabetian,C.P., Goldwurm,S. *et al*. (2008) Phenotype, genotype, and worldwide genetic penetrance of LRRK2-associated Parkinson's disease: a case-control study. *Lancet Neurol.*, **7**, 583–590.
21. Dendrou,C.A., Plagnol,V., Fung,E., Yang,J.H., Downes,K., Cooper,J.D., Nutland,S., Coleman,G., Himsworth,M., Hardy,M. *et al*. (2009) Cell-specific protein phenotypes for the autoimmune locus IL2RA using a genotype-selectable human bioresource. *Nat. Genet.*, **41**, 1011–1015.
22. Armitage,P., Berry,G. and Matthews,J.N.S. (2002) *Statistical Methods in Medical Research*. Blackwell Publishing Company, Malden, MA, USA.