International Journal of Antimicrobial Agents 53 (2019) 197–202

Contents lists available at ScienceDirect

International Journal of Antimicrobial Agents

journal homepage: www.elsevier.com/locate/ijantimicag

Hot Topic

Groups of coevolving positions provide drug resistance to *Mycobacterium tuberculosis*: A study using targets of first-line antituberculosis drugs

Sharad Vats[a], Asheesh Shanker[b,*]

[a] Department of Bioscience and Biotechnology, Banasthali Vidyapith, Rajasthan, India
[b] Department of Bioinformatics, Central University of South Bihar, Gaya, Bihar, India

ARTICLE INFO

Article history:
Received 5 July 2018
Accepted 20 October 2018

Editor: Jean-Marc Rolain

Keywords:
Mycobacterium
Drug resistance
Coevolution
Bioinformatics

ABSTRACT

Drug resistance has been associated with point mutations in coding regions leading to an altered protein sequence and structure. Such changes have been seen as isolated events occurring at various positions in a sequence. However, we hypothesise that it is not a single mutation at a specific position but a group of positions that coevolve in a correlated fashion to increase the fitness of a target protein against a drug. To prove the hypothesis, selected protein sequences of *Mycobacterium tuberculosis* drug resistance genes were successfully screened using a bioinformatics approach to detect groups of coevolving amino acids at important structural and functional positions in the targets of first-line antituberculosis drugs (isoniazid, rifampicin, ethambutol and pyrazinamide). The algorithmically characterised genetic mutations and the lineage-specific single nucleotide polymorphisms (SNPs) detected previously in drug resistance genes of *M. tuberculosis* complex genomes were also found in the identified coevolving groups. Mapping of coevolving positions to the secondary structure of proteins clearly indicates the preference of amino acid residues in the helix to coevolve. Moreover, active-site residues of some candidate proteins were also found in coevolving groups. The coevolving groups detected in this study will be useful to gain new insights into the molecular and evolutionary basis of drug resistance. This work provides an important first step towards finding solutions to the multidrug resistance problem through coevolution analysis of proteins, in turn helping to develop new drug regimens against pathogens, including *M. tuberculosis*.

1. Introduction

It is well recognised that the primary structure of a protein governs its folding in the three-dimensional (3D) structure, which in turn dictates its function. Since several amino acids in the polypeptide chain help to maintain the 3D structure, consequently they experience structural and/or functional constraints [1,2]. Due to this, most of the amino acids in proteins may not evolve independently but instead coevolve in a complementary fashion [3]. Coevolution occurs to fix an evolutionary destabilising change at one position by a compensatory change nearby. However, beneficial mutations are rarely found in combination [4]. Moreover, it is not necessary that all coevolving amino acids maintain direct physical contact [5]. Identification of coevolving amino acids in a protein has biological importance. It helps to know the structural and/or functional constraints acting on specific amino acids of a protein molecule. Such information may help to understand the mechanism of molecular evolution and also play an important role in structure prediction, protein engineering and annotation [3,5].

Nowadays, drug resistance due to point mutations is a common problem. Here we hypothesise that it is not a single mutation but instead a group of positions that coevolve in a correlated fashion to increase the fitness of a target protein against a drug. Previous studies on *Mycobacterium tuberculosis* have been conducted considering resistance-conferring mutations in nucleotide sequences [6,7], however evolutionary changes in proteins have been ignored. Moreover, owing to their complex mechanism, very few compensatory mutations were observed in *M. tuberculosis* [8–10]. The present study was therefore undertaken to test the hypothesis of coevolution in conferring resistance to pathogens using protein sequences that are widely used as the targets of first-line antituberculosis drugs.

* Corresponding author. Tel.: +91 94 1447 8655.
E-mail address: ashomics@gmail.com (A. Shanker).

https://doi.org/10.1016/j.ijantimicag.2018.10.027
0924-8579/© 2018 Elsevier B.V. and International Society of Chemotherapy. All rights reserved.

Table 1
Information on candidate sequences used in the study.

Drug	Gene	UniProt ID	No. of homologous sequences
Isoniazid	*katG*	P9WIE5	176
	inhA	P9WGR0	124
	folA	P9WNX1	120
	ndh	P95160	151
	ahpC	P9WQB7	105
	fabG1	P9WGT3	117
Rifampicin	*rpoB*	P9WGY9	186
Ethambutol	*embC*	P9WNL5	391
	embB	P9WNL6	389
	embA	P9WNL8	388
	embR	P9WGJ9	130
Pyrazinamide	*pncA*	Q50575	145
	rpsA	P9WH43	128

2. Materials and methods

2.1. Selection of candidate sequences

The scientific literature was searched to identify drug resistance genes (Table 1) relevant to the current first-line multidrug regimen (isoniazid, rifampicin, ethambutol and pyrazinamide) used to treat tuberculosis caused by *M. tuberculosis*. DrugBank (www.drugbank.ca) and Mycobrowser (https://mycobrowser.epfl.ch/) databases were used to acquire the UniProt ID of proteins corresponding to drug resistance genes. Basic Local Alignment Search Tool (BLAST) [11] was used to generate the data set of homologous sequences. Sequences that do not belong to *Mycobacterium*, putative, uncharacterised, fragments and those with ambiguous characters were removed.

2.2. Multiple sequence alignment and reconstruction of phylogenetic trees

MAFFT (http://mafft.cbrc.jp/alignment/server/) [12] with default parameters was used to generate multiple sequence alignment (MSA) of protein sequences. PhyML [13] with an improved general amino acid replacement matrix LG [14] was used to generate the phylogenetic tree of each data set. The inferred trees were used for coevolution analysis.

2.3. Detection of coevolving positions

Coevolving residues were detected using clustering and compensation analyses as implemented in CoMap v.1.5.2 [3,15]. This program effectively identifies co-substitutions, non-independent sites undergoing substitution in the same branches of the tree [16]. CoMap uses a set of aligned sequences, their phylogenetic tree, a substitution model and discrete rate distribution across sites to map site-specific substitutions along the tree. Biochemical properties including charge, polarity, volume and Grantham score were used to provide weights to the substitutions of amino acids. In addition, unweighted substitutions were also detected. An R-program also available with CoMap was used with 1000 replicates of parametric bootstrap analysis to check the statistical significance ($P \leq 0.05$) of identified coevolving sites and to assess the false discovery rate. Circos [17] was used to depict the coevolving positions in a circular plot. A mask for each amino acid residue in MSA was generated (Table 2) [18,19] to determine the residue class of coevolving amino acids.

Table 2
Amino acid residue masks based on residue class.

Residue class	Mask	Amino acids
Acid	A	D, E
Base	B	R, K
Non-aromatic hydrophobic	N	A, L, I, V, M
Non-charged polar	Q	S, T, C, N, Q
Aromatic	R	F, Y, W, H
Proline	P	P
Glycine	G	G

2.4. Active site and secondary structure analysis of coevolving residues

Protein Data Bank (PDB) files were used to look for secondary structure information (α helix or β-sheet) of coevolving amino acids. PDB files were available only for P9WIE5 (PDB ID: 1SJ2), P9WGR0 (4TRJ), P9WGY9 (4KBM), P9WNL5 (3PTY) and Q50575 (3PL1). Only chain A of PDB files was used to retrieve secondary structure data. Moreover, the SITE record of PDB files was used to recognise active-site residues. However, 4KBM does not contain the SITE record. The identified coevolving amino acids were mapped on the secondary structure and active-site residues.

2.5. Mapping of coevolving positions with known mutations

In the recent past, genetic mutations were algorithmically characterised as not conferring resistance (benign), lineage-defining, uncharacterised and resistance determinants for all first- and second-line drugs for 2099 *M. tuberculosis* genomes [7]. Moreover, Coll et al. [6] investigated lineage-specific single nucleotide polymorphisms (SNPs) at drug resistance genes in a global collection of 1601 *M. tuberculosis* complex (MTC) genomes. An attempt was made to map the identified coevolving positions on these mutations.

3. Results and discussion

It has previously been reported that in bacterial populations alleles related to drug resistance impose a fitness cost [20–22] and compensatory evolution may help in the stability of such mutants. These studies were based on analysis of nucleotide sequence data. Therefore, in the present analysis we identified groups of coevolving positions in protein sequences that are used as targets of first-line drugs against *M. tuberculosis*.

3.1. Identification of coevolving positions

Coevolving groups detected both by compensation (total 555; charge 64, Grantham score 171, polarity 138 and volume 182) and correlation (total 455; charge 32, Grantham score 83, polarity 34, volume 69 and simple 237) analysis in the drug targets against *M. tuberculosis* are shown in Fig. 1. It is evident from this figure that compensation analysis detects more coevolving groups than correlation analysis. However, correlation analysis detects a larger coevolving group size in most of the drug targets considered than compensation analysis (Fig. 2A). A higher number of unique coevolving positions was detected in compensation analysis (Fig. 2B). Moreover, individual positions involved in more than one coevolving group were found to be almost similar in both analyses (Fig. 2C). Compensation analysis detected more coevolving groups that are common among different biochemical properties (charge, polarity, volume and Grantham score). However, very few common coevolving groups were detected between compensation and correlation analysis (Supplementary Table S1). All of the coevolving groups detected are given in the supplementary

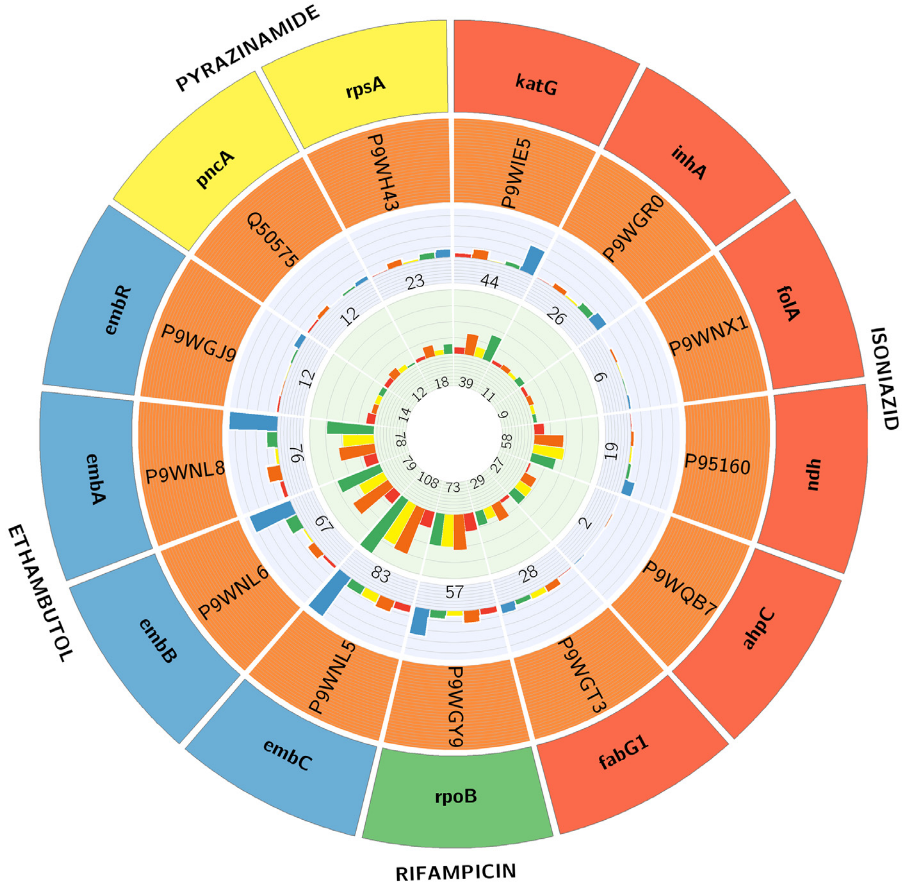

Fig. 1. Coevolving groups detected in the selected drug targets of *Mycobacterium tuberculosis*. The outer circle represents the candidate genes related to the current first-line multidrug regimen of isoniazid (6 genes), rifampicin (1 gene), ethambutol (4 genes) and pyrazinamide (2 genes). The orange circle shows the UniProt ID/Swiss-Prot accession no. of the respective protein sequence. The light blue and light green circles depict the coevolving groups detected in compensation and correlation analysis, respectively. The coloured bars show the coevolving groups detected by unweighted (simple; blue) and weighted substitutions based on different amino acid properties: charge (red); Grantham score (orange); polarity (yellow); and volume (green).

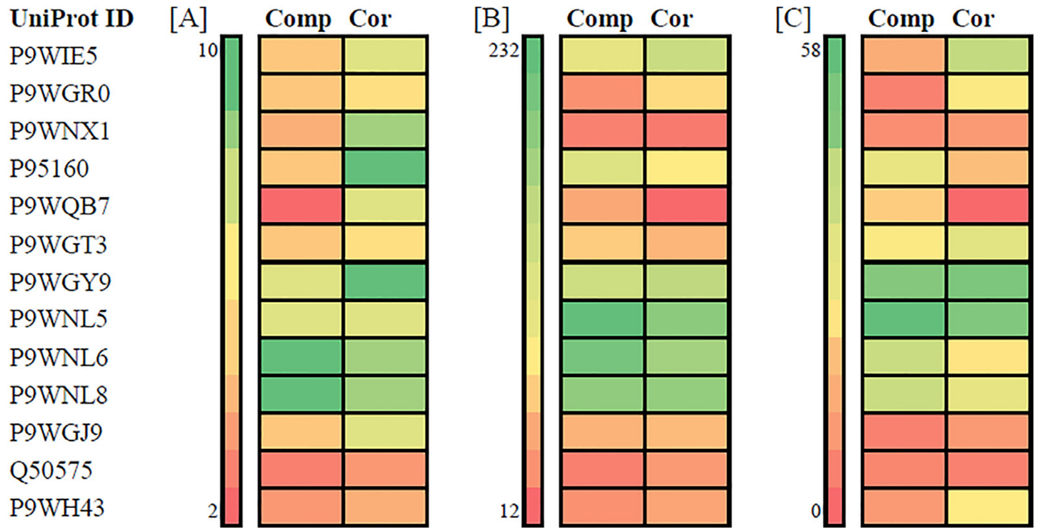

Fig. 2. (A) Size of coevolving groups, (B) unique coevolving positions detected and (C) coevolving positions involved in more than one group in the drug targets of *Mycobacterium*. The UniProt ID represents the drug targets considered. Comp, compensation; Cor, correlation.

files. Some of the coevolving groups detected by compensation analysis in the *katG* gene (P9WIE5), a target of the drug isoniazid, along with their amino acid residue class (in bold) are represented in Table 3. All coevolving groups detected by correlation and compensation analysis in *katG* along with their amino acid residue class are presented in Fig. 3 and Supplementary Fig. S1, respectively.

Recently, Singh et al. studied the effect of G279D mutation on the functionality of the KatG protein through structural analyses and reported that the mutant protein showed a smaller binding

Table 3
Some of the coevolving groups detected by compensation analysis in the *katG* gene (Swiss-Prot accession no. P9WIE5) along with their amino acid residue class (in bold).

Coevolving group	Size	*P*-value	Weight [a]	FDR
A/D/E 241; D/G/N/S 477; A/G/R 277; E/G 677; E/I/V 284; A/E 433 **A/N 241; A/G/Q 477; B/G/N 277; A/G 677; A/N 284; A/N 433**	6	0.006 **	Charge	Yes
A/G/R/S/T 48; D/V 283; E/G/I/T/V 415; A/G/R 625; A/R/S 720 **B/G/N/Q 48; A/N 283; A/G/N/Q 415; B/G/N 625; B/N/Q 720**	5	0.023 *	Charge	Yes
A/D/K/N/S/T 500; A/D/G/N/S 561; A/K/L/P/S/V 676 **A/B/N/Q 500; A/G/N/Q 561; B/N/P/Q 676**	3	0.019 *	Charge	Yes
S/T/V 247; I/M/N/V 309 **N/Q 247; N/Q 309**	2	0.017 *	Grantham score	Yes
I/M/V 338; I/L 484; S/T 512; I/T/V 703; F/H/Y 664; D/E 742 **N 338; N 484; Q 512; N/Q 703; R 664; A 742**	6	0.036 *	Grantham score	Yes
A/P 380; A/P 392 **N/P 380; N/P 392**	2	0.005 **	Grantham score	Yes
A/D/E 241; D/G/N/S 477 **A/N 241; A/G/Q 477**	2	0.008 **	Polarity	Yes
A/P 380; A/P 392 **N/P 380; N/P 392**	2	0.006 **	Polarity	Yes
E/L/Q/W 462; E/I/M/Q/V 730 **A/N/Q/R 462; A/N/Q 730**	2	0.003 **	Polarity	Yes
A/P 380; A/P 392 **N/P 380; N/P 392**	2	0.006 **	Volume	Yes
P/Y 41; A/I/K/L/R/V 529 **P/R 41; B/N 529**	2	0.004 **	Volume	Yes
A/G/I/S 69; A/L/R 489; E/F/L/W 456; A/R/S 720 **G/N/Q 69; B/N 489; A/N/R 456; B/N/Q 720**	4	0.022 *	Volume	Yes

FDR, false discovery rate.
[a] Biochemical property used for coevolution analysis.* Significant; ** very significant.

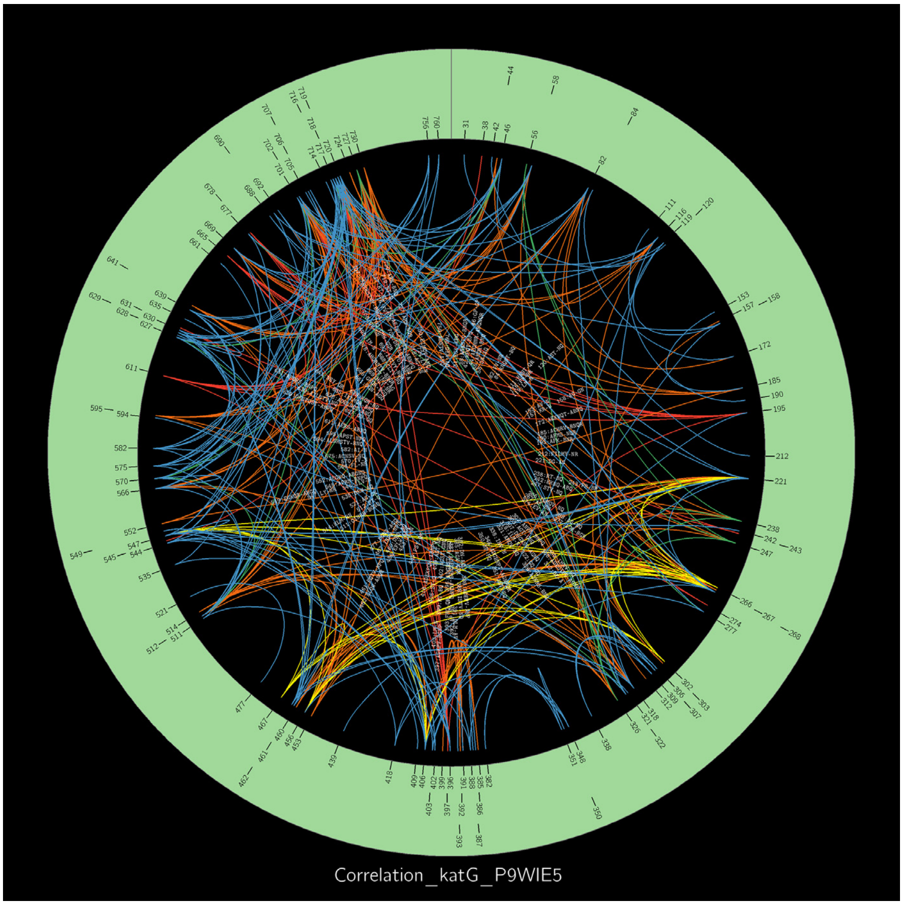

Fig. 3. Coevolving groups detected by correlation analysis in the *katG* gene (Swiss-Prot accession no. P9WIE5). Links show the coevolving groups detected based on simple (blue), charge (red), Grantham (orange), polarity (yellow) and volume (Green) analysis. Coevolving positions are shown within the thick light green circle. The inner alphanumerical circle represents coevolving positions along with the amino acid residue and their residue class (coevolving position: amino acid residue–amino acid residue class).

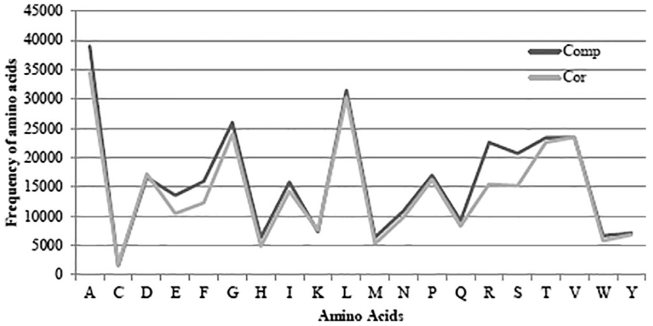

Fig. 4. Frequency of coevolving amino acids detected both in compensation (Comp) and correlation (Cor) analysis. Each coevolving position is considered only once to avoid over-representation of an amino acid. Amino acids are represented as a one letter code and are arranged alphabetically.

cavity, lower docking score and reduced affinity towards isoniazid [23]. However, as per our study, it seems important to consider mutations that act in a group along with the resistance-causing mutations when designing new drugs. Therefore, instead of exploring the effect of a single amino acid change on protein structure, the combined effect of groups of coevolving positions must be considered. Although several resistance-causing mutations are known [6,7], our analysis demonstrated that there is a need to explore all complementary mutations in the form of coevolving groups to combat pathogens.

3.2. Frequency of coevolving amino acids

The frequency of coevolving amino acids detected both in compensation and correlation analysis is presented in Fig. 4. To avoid over-representation of an amino acid, each coevolving position was taken only once to calculate the frequency. An independent sample *t*-test suggested that on average the frequency of coevolving amino acids detected both in compensation and correlation analysis is homogeneous, i.e. no significant difference between them. Considering compensation and correlation analyses, alanine (12.09%) contributed most to the coevolving amino acids whilst cysteine (0.55%) was least abundant.

The amino acid residues alanine, glycine and valine ranked 1, 2 and 4 based on their abundance in coevolving positions and are among the five amino acid residues reported by Holliday et al. whose side chains are never involved in catalysis in MACiE (a database of enzyme reaction mechanisms) [24]. This suggests that these residues are not important for the functioning of proteins and therefore are more often involved in coevolution.

The distribution of amino acid residues involved in coevolution considering their residue class is presented in Supplementary Fig. S2. Most of the coevolving positions contain non-aromatic hydrophobic amino acids (residue mask **N**: A, L, I, V, M; 36.86%), followed by non-charged polar (**Q**: S, T, C, N, Q; 20.33%), aromatic (**R**: F, Y, W, H; 10.87%), acidic (**A**: D, E; 9.53%), basic (**B**: R, K; 8.72%), glycine (**G**; 8.22%) and proline (**P**; 5.47%).

3.3. Secondary structure and active-site analysis

Mapping of coevolving positions to the secondary structure clearly indicates the preference of amino acid residues in the helix to coevolve. These results are consistent with earlier findings which suggest that coevolution mostly affects the helical parts of the structure [25–27]. Moreover, it indicates that coevolution has a direct impact on the structural integrity of a protein.

The active-site residues of 1SJ2, 4TRJ and 3PTY were also found in coevolving groups (Supplementary Table S2). The results reveal that coevolution can modulate the active-site residues thereby having an effect on the function of a protein. The active-site residue Trp985 (PDB ID 3PTY/*embC*/P9WNL5) is one of the amino acids found as coevolving. Alderwick et al. also reported that a single-residue substitution of conserved tryptophan residues (Trp868, Trp985) at these respective sites inhibited EmbC-catalysed extension of lipoarabinomannan found in the unique cell envelope of *M. tuberculosis* [28]. However, changes in amino acids distant from the active site may also be functionally important [2], and it was also observed in directed evolution experiments that beneficial mutations often occur where least expected [29]. Thus, all of the mutations in a coevolving group may not be directly involved in providing resistance, but considering them during the detection/prediction of resistance-determining residues and related structural analysis could be beneficial.

3.4. Known drug resistance mutations and coevolving positions

The algorithmically characterised benign, lineage-defining, resistance determinants and uncharacterised genetic mutations [7] and the lineage-specific SNPs [6] detected in drug resistance genes of MTC genomes were also involved in coevolving groups identified in this study (Supplementary Table S3). As indicated by the findings of the present analysis, it is possible that the development of resistance requires substitution at various positions. Therefore, we suggest that groups of coevolving positions in proteins may also be considered along with resistance-determining mutations, which will help in the designing of more efficient drugs against pathogens.

Acknowledgments

The authors are grateful to Prof. Nicolas Salamin (University of Lausanne, Lausanne, Switzerland) and Dr Julien Y. Dutheil (Max Planck Institute for Evolutionary Biology, Plön, Germany) for generous help in comap execution. The authors are also grateful to the anonymous reviewer for useful suggestions to improve the manuscript.

Funding

None.

Competing interests

None declared.

Ethical approval

Not required.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ijantimicag.2018.10.027.

References

[1] Abriata LA, Palzkill T, Dal Peraro M. How structural and physicochemical determinants shape sequence constraints in a functional enzyme. PLoS One 2015;10:e0118684. doi:10.1371/journal.pone.0118684.
[2] Jack BR, Meyer AG, Echave J, Wilke CO. Functional sites induce long-range evolutionary constraints in enzymes. PLoS Biol 2016;14:e1002452. doi:10.1371/journal.pbio.1002452.
[3] Dutheil J, Pupko T, Jean-Marie A, Galtier N. A model-based approach for detecting coevolving positions in a molecule. Mol Biol Evol 2005;22:1919–28.
[4] Arnold FH. Design by directed evolution. Acc Chem Res 1998;31:125–31.

[5] Pazos F, Valencia A. Protein co-evolution, co-adaptation and interactions. EMBO J 2008;27:2648–55.

[6] Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, et al. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. Nat Commun 2014;5:4812.

[7] Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, Bradley P, et al. Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. Lancet Infect Dis 2015;15:1193–202 Erratum in: Lancet Infect Dis 2018;18:21.

[8] Sherman DR, Mdluli K, Hickey MJ, Arain TM, Morris SL, Barry CE 3rd, et al. Compensatory *ahpC* gene expression in isoniazid-resistant *Mycobacterium tuberculosis*. Science 1996;272:1641–3.

[9] Shcherbakov D, Akbergenov R, Matt T, Sander P, Andersson DI, Böttger EC. Directed mutagenesis of *Mycobacterium smegmatis* 16S rRNA to reconstruct the in-vivo evolution of aminoglycoside resistance in *Mycobacterium tuberculosis*. Mol Microbiol 2010;77:830–40.

[10] Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, et al. Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. Nat Genet 2011;44:106–10.

[11] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.

[12] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 2013;30:772–80.

[13] Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 2010;59:307–21.

[14] Le SQ, Gascuel O. An improved general amino acid replacement matrix. Mol Biol Evol 2008;25:1307–20.

[15] Dutheil J, Galtier N. Detecting groups of coevolving positions in a molecule: a clustering approach. BMC Evol Biol 2007;7:242.

[16] Tuffery P, Darlu P. Exploring a phylogenetic approach for the detection of correlated substitutions in proteins. Mol Biol Evol 2000;17:1753–9.

[17] Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res 2009;19:1639–45.

[18] Hecht D, Tran J, Fogel GB. Structural-based analysis of dihydrofolate reductase evolution. Mol Phylogenet Evol 2011;61:212–30.

[19] Shanker A. Structural analysis of form I ribulose-1, 5-bisphosphate carboxylase/oxygenase. BAOJ Bioinfo 2016;1:001.

[20] Sander P, Springer B, Prammananan T, Sturmfels A, Kappler M, Pletschette M, et al. Fitness cost of chromosomal drug resistance-conferring mutations. Antimicrob Agents Chemother 2002;46:1204–11.

[21] Post FA, Willcox PA, Mathema B, Steyn LM, Shean K, Ramaswamy SV, et al. Genetic polymorphism in *Mycobacterium tuberculosis* isolates from patients with chronic multidrug-resistant tuberculosis. J Infect Dis 2004;190:99–106.

[22] Gagneux S, Long CD, Small PM, Van T, Schoolnik GK, Bohannan BJ. The competitive cost of antibiotic resistance in *Mycobacterium tuberculosis*. Science 2006;312:1944–6.

[23] Singh A, Singh A, Grover S, Pandey B, Kumari A, Grover A. Wild-type catalase peroxidase vs G279D mutant type: molecular basis of isoniazid drug resistance in *Mycobacterium tuberculosis*. Gene 2018;641:226–34.

[24] Holliday GL, Almonacid DE, Mitchell JB, Thornton JM. The chemistry of protein catalysis. J Mol Biol 2007;372:1261–77.

[25] Pollock DD, Taylor WR, Goldman N. Coevolving protein residues: maximum likelihood identification and relationship to structure. J Mol Biol 1999;287:187–98.

[26] Dimmic MW, Hubisz MJ, Bustamante CD, Nielsen R. Detecting coevolving amino acid sites using Bayesian mutational mapping. Bioinformatics 2005;21(Suppl 1):i126–35.

[27] Wang M, Kapralov MV, Anisimova M. Coevolution of amino acid residues in the key photosynthetic enzyme Rubisco. BMC Evol Biol 2011;11:266.

[28] Alderwick LJ, Lloyd GS, Ghadbane H, May JW, Bhatt A, Eggeling L, et al. The C-terminal domain of the arabinosyltransferase *Mycobacterium tuberculosis* EmbC is a lectin-like carbohydrate binding module. PLoS Pathog 2011;7:e1001299.

[29] Farinas ET, Bulter T, Arnold FH. Directed enzyme evolution. Curr Opin Biotechnol 2001;12:545–51.