An evolution-based method of epistasis measurement: theory and application to influenza

Gabriele Pedruzzi and Igor M. Rouzine*

Sorbonne Université, Institute de Biologie Paris-Seine
Laboratoire de Biologie Computationelle et Quantitative LCQB
F-75004, Paris, France

Influenza protein sequence data are from public database https://www.fludb.org.

Short Title: Epistasis measurement

Key words: mathematical model, quasi-equilibrium, false-positive, linkage

***Corresponding author:**
Igor Rouzine
igor.rouzine@sorbonne-universite.fr
Equipe "Evolution and immunology of pathogens"
Laboratory of Computational and Quantitative Biology
7238 CNRS-UPMC
Bâtiment C, C311C
Sorbonne Université Campus Pierre et Marie Curie
4 rue Jussieu, Paris 75005
+33 7 69 54 91 28
http://www.ibps.upmc.fr/fr/Recherche/umr-7238/evolution-immunologie-pathogenes

Abstract

Linkage effects in a multi-locus population strongly influence its evolution. Recent models based on the traveling wave approach enable us to predict the speed of evolution and the statistics of phylogeny. However, predicting the evolution of specific sites and pairs of sites in the multi-locus context remains a mathematical challenge. In particular, the effect of epistasis, the interaction of gene regions contributing to phenotype, is difficult both to predict theoretically and detect experimentally in sequence data. A large number of false interactions arising from linkage and indirect interactions which mask true interactions. Here we develop a method to clean false-positive interactions. We start by demonstrating that averaging of the two-way haplotype frequencies over a hundred of independent populations is not enough to clear false interactions. Then, to address this problem, we develop analytically and use a triple-way haplotype test, which isolates true interactions. Next, the fidelity of the test is confirmed on simulated genetic sequences, where the epistatic network known in advance. Finally, we apply the test to a large database on influenza A H1N1 virus sequences of neurominidase from various geographic locations to predict the epistatic network responsible for the transition from the pre-pandemic virus to the pandemic strain. We predict a primary mutation and 15-22 secondary compensatory mutations of variable strength, as many as typically observed for drug resistance and immune escape mutations in HIV. These results present a simple and reliable method to measure epistatic interaction from sequence data.

Author's summary

Interaction of genomic sites creating "fitness landscape" is very important for predicting the escape of viruses from drugs and immune response and for pssinf through fitness valleys. Many efforts have been invested into measuring these interactions from DNA sequence sets. Unfortunately, reproducibility of the results remains low, due partly to a very small fraction of interaction pairs, and partly to stochastic noise intrinsic for evolution masking true interactions. Here we propose a method based on analysis of genetic sequences at three genomic sites to clean stochastic linkage and apply it to influenza virus sequence data.

Introduction

It was realized long ago that properties of an evolving populations are strongly affected by the fact that alleles at different genetic sites, loci, are co-inherited unless separated by recombination. These *linkage* effects include clonal interference (FISHER 1930; MULLER 1932), genetic background and genetic hitchhiking effect (RICE 2004), enhanced accumulation of deleterious mutations (Muller ratchet) (FELSENSTEIN 1974), and the increase of random genetic drift on one locus due to selection at another locus (HILL AND ROBERTSON 1966a). Linkage slows down substantially adaptation of population, and creates random associations between pairs of mutations that occurred on the same branch of the ancestral tree.

These effects has been taken into account by early mathematical models considering two loci (KIMURA 1994) and, more recently, by the traveling wave approach, which can describe an arbitrarily large number of connected sites (ROUZINE *et al.* 2003; DESAI AND FISHER 2007; HALLATSCHEK 2011; GOOD *et al.* 2012). These models usually consider fitness classes of sequences subjected to selection, mutation, random genetic drift. They show that population represent a narrow fitness distribution traveling in the fitness space in a direction depending on initial

conditions and parameters. It consists from the deterministic bulk subject to selection alone, and the stochastic leading edge. At the edge, generation and establishment of new beneficial mutations, limits the speed of adaptation. Alternatively, the distribution may move backwards accumulating deleterious alleles (Muller ratchet). These models were able to predict accurately several integral parameters of evolution in terms of model parameters, including population size, mutation rate, and the distribution of selection coefficients over loci. Predicted parameters include the adaptation rate, Muller ratchet rate, equilibrium state (ROUZINE *et al.* 2003; ROUZINE *et al.* 2008), the probability of fixation and the probable selection coefficient (GOOD *et al.* 2012), and the statistics of the ancestral tree (WALCZAK *et al.* 2012; NEHER AND HALLATSCHEK 2013).

With all this progress, the evolution of separate sites in the multi-site still remains an open question. How do allelic frequencies at each site change in time? Although they are known to follow random trajectories, what can be said about the average allelic frequency of a given site with a given fitness effect of mutation? What can we say about the evolution of site pairs, especially, in the presence of the epistatic interaction?

Epistasis, the interaction of genes and gene regions contributing to phenotype, is an omnipresent phenomenon (CARLBORG *et al.* 2006). Gene interactions are responsible for 70% of the organism's genetic inheritance (ZUK *et al.* 2012) and create fitness valleys in the evolutionary path (WEISSMAN *et al.* 2009). In pathogens, epistasis accelerates the development of drug resistance and immune escape and impedes its reversion (NIJHUIS *et al.* 1999; LEVIN *et al.* 2000; PIANA *et al.* 2002; HANDEL *et al.* 2006; GONZALEZ-ORTEGA *et al.* 2011; WU *et al.* 2018). Most of HIV variation in untreated patients has been argued to arise from mutations compensating virus escape from the immune response (ROUZINE AND COFFIN 1999).

A number of approaches have been proposed to measure epistasis from genomic data (CORDELL 2009; CHEN *et al.* 2011; UEKI AND CORDELL 2012). The simplest methods are based on pairwise allelic correlations (HILL AND ROBERTSON 1966b; BARTON 1995). However, other forces, such as linkage and indirect interactions, create strong correlations even between non-interacting pairs of loci, much more

numerous than epistatic pairs. Stochastic effects are a serious obstacle to the detection of epistatic effects (WEI *et al.* 2014). Mathematical modeling shows that, in a single asexual population with a relative short genome, stochastic linkage overshadows the epistatic footprint, except in a narrow range of times and parameters (PEDRUZZI AND ROUZINE 2019).

A tree-based real-time technique was used to detect epistatic interactions in influenza A including those that are known to confer resistance to drug oseltamavir (KRYAZHIMSKIY *et al.* 2011; NEVEROV *et al.* 2015). They identified compensatory mutations as closely following primary mutations and appearing in the same lineage. A method to discern directly interacting pairs from false (indirect) interactions in a highly-diverged single-time protein sequence set from different species has also been proposed (WEIGT *et al.* 2009; COCCO *et al.* 2018). The idea was to combine a covariance method based on mutual information with the estimates of the direct mutual information. The direct interactions are estimated from the condition that entropy is maximal given observed frequencies of single sites and pairs of sites. A similar technique has been used to measure the fitness landscape of Ab binding regions of HIV surface protein gp120 (LOUIE *et al.* 2018).

While the last data-based technique allows to eliminate indirect interactions for highly-divergent sequences from different species, it cannot eliminate linkage effects present in a single or several recently-diverged populations. The divergence of independent populations from a common ancestor in a random direction creates high covariance (linkage disequilibrium) of random sign for most of site pairs masking the covariance from epistasis (PEDRUZZI AND ROUZINE 2019).

Recently, we proposed another, evolution-based approach to the problem. The challenge we faced was to predict analytically the co-evolution of a specific pair of loci linked to a multi-site genome and then use this information to estimate epistatic interaction (PEDRUZZI *et al.* 2018). Using a simulated population, we verified that, in a broad parameter range, the distribution of alleles across genomes and sites, within each fitness class, is controlled by the condition of maximum entropy at given fitness. We demonstrated the existence of a broad parameter and time range where each fitness class of an adapting population has sufficient time to

assume the most likely and disordered state, given by the current average fitness, which changes slowly in time. This situation is referred to as "quasi-equilibrium" (not to be confused with "linkage quasi-equilibrium" or "mutation-selection balance") . In quasi-equilibrium, the entropy S is at its current maximum and is a function of fitness W, which slowly changes in time during the process of adaptation: $S = F[W(t)]$. The following simple derivation employs the fact, known from all the traveling-wave models cited above, that the distribution of fitness W is relatively narrow, $\mathrm{std}[W] \ll W$. Quasi-equilibrium is possible due to a very slow motion of the fitness distribution, limited by the slow establishment of new beneficial alleles at the front edge of the distribution (ROUZINE *et al.* 2003; DESAI AND FISHER 2007; NEHER *et al.* 2010; HALLATSCHEK 2011). This situation exists very far mutation-selection balance, i.e., full equilibrium.

Consider a pair of loci (sites). Assume that each site can have two variants, either a better-fit allele or a less-fit allele, denoted 0 or 1, and s_1, s_2 are the respective fitness effects of mutation at the sites. We can have 4 possible haplotypes, 00, 01, 10, 11. Let us also assume that difference in fitness between two alleles is the two sites interact epistatically, so that two deleterious alleles emerging together, denoted 11, partly compensate each other to the degree E, where $0 < E < 1$. The log fitness values of the four haplotypes are

$$W_{00} = 0,\ W_{01} = -s_1, W_{10} = -s_2,\ W_{11} = -(s_1 + s_2)(1 - E),$$

respectively (Fig. 1). If E is allowed to change below 0 or above 1, we can obtain any type and sign of epistasis: positive, negative, overcompensation.

The frequency of haplotype in population f_{ij} is proportional to the number of possible sequence configurations of the rest of the genome (grey box in Fig. 1a), $\exp(S_{\mathrm{rest}})$. Entropy S_{rest}, as we explained above, is restricted by the fitness of the rest of genome equal to $W - W_{ij}$. Hence, the entropy of each haplotype subset is $S_{\mathrm{rest}} = S(W - W_{ij})$. Since the genome is long, we can safely assume that W_{pair} is much smaller

than W, so that the corresponding change in entropy is small and proportional to W_{ij}. Hence, we can approximate

$$S(W - W_{ij}) \approx S(W) - \beta W_{ij} \tag{1}$$

where $\beta = |dS/dW|$. The frequency of each haplotype is proportional to the corresponding configuration number, $\exp[S(W - W_{ij})]$. Combining Eqs. 1 and 2, we can express the haplotype frequencies as

$$\begin{aligned} f_{10} &= f_{00}e^{-\beta s_1} \\ f_{01} &= f_{00}e^{-\beta s_2} \\ f_{11} &= f_{00}e^{-\beta(s_1+s_2)(1-E)} \end{aligned} \tag{2}$$

where β slowly depends on W and, hence, on time, but is the same for all genomic sites (PEDRUZZI *et al.* 2018). This simple derivation assumes that the site pair does not interact with other sites of the genome.

This assumption can be lifted and generalized for any interaction network, and the function $S(W)$ can be calculated explicitly from a more microsopic model (PEDRUZZI *et al.* 2018). We also do not need to restrict ourselves to the bi-allelic model, but we will keep this assumption, since it is fairly good for a relatively short term evolution. (A physicist will recognize in Eq. 2 the Boltzman distribution derived originally for a small subsystem connecting to a large thermal reservoir.)

Eqs. 1, 2 predicts how the average frequencies of single site and pairs of sites evolve in time. It is also tempting to use it for the practical goal of estimating parameters E, s_1, s_1 from the observed haplotype frequencies, and thus find the fitness landscape for the entire genome, at least, if epistasis is pairwise, as we assumed. In reality, these formulae are not readily applicable to real-life sequences. because of the strong random variation of f_{ij} among pairs of loci caused by stochastic linkage effects.

Of course, the problem of strong linkage noise causing poor reproducibility of various methods [see (WEI *et al.* 2014) for review] is not specific for this method of estimating epistasis. Any attempt to measure epistasis in a single population, whether by measuring covariance [D', r^2, mutual entropy, (PEDRUZZI AND ROUZINE 2019)] or by tree-based methods used, for example, by (NEVEROV *et al.* 2015), faces the same problem. The epistatic pairs in a single population are not observable in principle, due to the the phylogenetic relationship of sequences in population. As shown previously by simulation, all sequences in a population closely resemble the common ancestor, which diverges from the origin in a random direction (PEDRUZZI AND ROUZINE 2019). As a result, any measure of co-variance, or even the use of the entire tree, produces only strong noise of random sign. Co-variation due to random linkage completely masks the epistasis signature in a population. The only way to resolve this issue is by averaging over many independent populations. But how many populations do we need to have a small fraction of false positives?

As we show in the present work, the number of required populations is unrealistically large even for a short genome of 50 loci, not to mention more realistic genomes. The averaging over populations helps only to enrich the observed network in true bonds, but does not solve the problem: false-positive interactions dominate. We also not that the existing technique developed for measuring epistasis by comparing sequences from remote species (WEIGT *et al.* 2009; COCCO *et al.* 2018), does not help with this problem either, because it is designed for indirect interactions, not for filtering out linkage effects for closely-related populations.

Here we offer a new technique, based on the application of the quasi-equilibrium argument (above) to triple-way haplotypes. We show analytically and by simulation that a triple-way haplotype test represents an effective way to eliminate residual false links, including linkage and indirect interactions. We will demonstrate its high fidelity in a broad parameter range on the mock sequences obtained by computer simulation.

After training the method on simulated sequences, we apply this technique to real virus sequences from an adapting population. Influenza virus evolving in a population was demonstrated to obey the traveling wave theory with an effective

237 selection pressure caused by the accumulating memory cells (ROUZINE AND ROZHNOVA
238 2018; YAN *et al.* 2019). Therefore, it satisfies the same quasi-equilibrium criteria
239 (the evolution is slow and the distribution of fitness is narrow) and is amenable to
240 our method. Using 8000 influenza sequences obtained from various geographic
241 locations sequences, both before and after the pandemic of 2009, we detect and
242 estimate quantitatively the epistatic network responsible for the transition of two
243 influenza proteins (NA and HA) between the old and new strains, which share 80%
244 of homology. To avoid a possible misunderstanding, the origin and the cause of
245 pandemics, such as antigenic shifts, are not addressed in this paper; we focus only
246 on the epistatic network, which allows to obtain the new variant from the old
247 variant.
248
249
250 **RESULTS**
251
252 **Simulation model to generate sequences for the test.** We start by simulating *in*
253 *silico* stochastic asexual evolution of a virus using a Wright-Fisher model in the
254 presence of the factors of mutation, random genetic drift, and selection with known
255 epistatic interactions (Fig. 2b). Simulation generates genome sequences
256 demonstrating LD originating from epistatic interactions, indirect interactions, and
257 stochastic linkage effects. We assume only two alleles per amino acid position (site):
258 better-fit (denoted 0) or less-fit (denoted 1). The binary simplification provides a
259 major reduction in the computational cost and is especially accurate for relatively
260 conserved sequences (ROUZINE AND COFFIN 2005). We set some pairs of sites to
261 interact epistatically, with a mutual degree of compensation between deleterious
262 alleles set at $E = 0.75$. We consider the same example of double arches (Fig. 1 or 2B).
263
264 **First step: averaging over populations.** Out task is to recover true epistatic
265 interactions from simulated sequences using pair-wise association analysis (Fig.
266 2A). Previous work has shown that their detection is masked by stochastic linkage,
267 which is known for its strong effects on evolution, including clonal interference,

genetic background effect, and alterations in the adaptation speed and genealogy (FISHER 1930; BARTON 1995; ROUZINE *et al.* 2003; BRUNET *et al.* 2008; ROUZINE AND COFFIN 2010; GOOD *et al.* 2012; NEHER AND HALLATSCHEK 2013). To decrease the linkage effects on detection, we calculate pairwise haplotypes frequencies f_{ij} for each pair of sites (*Methods*) and average them over multiple evolutionary-independent populations of the same size (PEDRUZZI AND ROUZINE 2019). Then, we perform a pairwise association analysis using a correlation measure shown to be a direct estimator of the degree of mutual compensation of two deleterious alleles, E (PEDRUZZI *et al.* 2018). More traditional LD measures, including D' and Pearson coefficient r^2, have similar performance but do not estimate E directly (PEDRUZZI AND ROUZINE 2019). We select pairs with high correlation, UFE > 0.6. For a single population, the raw cluster of predicted pairs is extremely complex and completely hides true epistatic interactions, in agreement with previous results (PEDRUZZI *et al.* 2018) (Fig 2C). A significant complexity reduction is obtained by averaging f_{ij} over 200 independent populations (Fig. 2D).

Second step: Triple-site haplotype method. Although much simpler, the resulting network in Fig. 1D is still densely populated by residual stochastic LD and indirect links, which are as strong as direct interactions. To filter the residual false links, we use a procedure based on the quasi-equilibrium approximation *(Introduction* and *Methods, Analytic Derivation*), and then test its validity by comparison of the predicted network with the network set in simulation, as follows.

We re-calculate UFE for every potential link *i,j* by considering only the sequences that contain 0 at one of the neighbor nodes of the link (3-locus haplotypes 110, 100, 010, and 000), and denote it UFE_{ij0}. A 0-node cuts a possible indirect correlation for the link even it is very strong (*Methods*). Considering different 0 nodes, we find the minimum value of UFE_{ij0} over all possible detours. The scatter plot in Fig. 2E demonstrates that, for the false pairs, $\min(\mathrm{UFE}_{ij0})$ is several-fold smaller than UFE_{ij}. The reason for this effect is that a 0-node removes a detour path around the link that causes indirect correlations. Therefore, we identify and remove false links as those with low $\mathrm{UFE}_{ij0}/\mathrm{UFE}$. The exact cutoff is not crucial, as

299 long as we average f_{ij} over, at least, 20 populations, so that the two groups, false and
300 true interactions, remain distinct. As a result, we obtain 100% perfect detection. We
301 even obtain correct estimates for the compensation strength, within 15% accuracy
302 UFE = E (Fig. 2F, green links).

303 We can conclude that, as predicted previously [2008, 2009], the quasi-
304 equilibrium approximation works in this parameter range, which corresponds to
305 the traveling wave regime, and that the triple-haplotype method has a high fidelity
306 with simulated sequences with a known epistatic network.

307

308 **Application to influenza A virus data.** After training the method on simulated
309 sequences, we calculate interaction network for virus sequences from an evolving
310 population. Influenza evolution in human population was shown to be described by
311 the traveling wave framework [rr, neher], which justifies the use of quasi-
312 equilibrium approximation during the process of adaptation (not to be confused
313 with mutation-selection equilibrium) [PBR]. Therefore, we have the right to apply
314 the technique to influenza virus sequences.

315 We focus on the two surface proteins of Influenza A virus strain H1N1,
316 Neuraminidase (NA) and Hemagglutinin (HA), both important targets of immune
317 response and drugs. Our aim is to identify subnetwork responsible for the transition
318 of the pre-2009 strain of H1N1 to the post-2009 of the same strain. For this aim, we
319 need to compare sequences of the first strain sampled worldwide to sequences of
320 the second strain sample worldwide.

321 We have downloaded 8440 and 9811 sequences for NA and HA, respectively,
322 from a public database (https://www.fludb.org). They were collected worldwide,
323 from different geographic locations, between years 2005 and 2010, and included
324 both pre-pandemic and post-pandemic strain. After aligning and binarizing
325 sequences by setting consensus residues to 0 and non-consensus to 1, we observed
326 a bimodal distribution of sequences in allele frequency, with two separate maxima.
327 The bimodal distribution is related to the emergence of a new strain with 80%
328 homology with the old strain in 2009.

329 In order to compensate for unequal sampling from the early strain and the
330 late strain, the sequences with mutation frequency per genome less than a preset
331 value d_v were randomly sampled and down-weighted by a coefficient D_w ranging
332 from 5% to 50%. To obtain the average pairwise haplotype frequencies f_{ij} , we
333 repeated the sampling multiple times. Then, we followed the procedure described
334 above to predict the intra-protein network of interactions.
335
336 **Sensitivity to sampling.** The resulting networks are fairly robust to the variation of
337 (d_v, D_w) (Fig. 3A-F). There exists a plateau region of (d_v, D_w) where the total number
338 of links varies weakly (Fig. 3E). The dependence of results on (d_v, D_w), which predict
339 between 15 and 22 compensatory sites, probably originates from the unequal
340 presentation of local subpopulations in the database. As a result, stochastic linkage
341 effects average out less than optimally. Below, we choose network variants shown in
342 Fig 3D and 3J as representative for NA and HA, respectively. We observe that site
343 248 in NA serves the primary site for multiple compensatory links (Fig. 3D).
344
345 **Structural interpretation.** The predicted epistatic sites are shown on the three-
346 dimensional protein structure of NA. The active pocket of NA (purple) serves to bind
347 sialic acid on target cell surface (Fig 4A). These results suggest that a mutation at
348 residue 248 near the pocket was the primary mutation responsible for the
349 replacement of the old variant of NA with the new variant of NA, and the other
350 mutations shown on the wheel are compensatory mutations helping NA to restore
351 and improve its fitness. Mutation 248 was shown to enhance the low-pH stability of
352 NA (TAKAHASHI *et al.* 2013), and it is found in all influenza A H1N1 strains after 2009
353 pandemic, regardless of a geographic location (SEIBERT *et al.* 2012; BYARUGABA *et al.*
354 2016; OTTE *et al.* 2016).
355
356
357 **DISCUSSION**
358

359 In the present work, we propose an efficient evolution-based method to identify the
360 co-variance caused by epistasis from the co-variance caused by stochastic linkage
361 effects and indirect interaction. The idea (*Methods*) is to pre-average the observed
362 haplotype frequencies over independent (or quasi-independent) populations, select
363 the links with high co-variance, and then apply a tri-way haplotype test for each
364 candidate bond to eliminate the effects of linked sites due to common inheritance.
365 The test is designed analytically ,based on the quasi-equilibrium state within
366 the traveling fitness wave (not to be confused with quasi-linkage equilibrium or
367 mutation-selection balance) tested by simulation in a broad parameter range
368 (PEDRUZZI *et al.* 2018; BARLUKOVA *et al.* 2020). Intuitively, because population
369 evolves slowly, the distribution of alleles between sites has sufficient time to obtain
370 the most likely and disordered state given a current average fitness of population.
371 We use a mock sequence set evolved in a Wright-Fisher population with a
372 known epistatic network to demonstrate the high fidelity of the method in the
373 controlled environment. The method eliminates false bonds caused by both indirect
374 interactions and linkage, at least, in the case of relatively simple epistatic network
375 and moderately diverse populations (< 40% diversity).
376 We apply this technique to identify primary and secondary mutations in the
377 two surface proteins of influenza A H1N1 responsible for the transition between the
378 pre-2009 variant and the post-2009 variants of these proteins. We do not address
379 the complex origins of 2009 pandemic in this work, which involves antigenic shifts.
380 Influenza virus has been shown to map to the traveling wave theory (ROUZINE AND
381 ROZHNOVA 2018; YAN *et al.* 2019), which justifies the use of the quasi-equlibrium
382 assumption. For NA protein, we observe a single primary site and 15-20 strong
383 multiple compensatory mutations, whose is in the same range as the number of
384 compensatory mutations observed for HIV for drug-resistant strains.
385 As compared to the existing technique of removal of the indirect bonds
386 developed for highly diverged species (WEIGT *et al.* 2009; COCCO *et al.* 2018), the
387 method is designed to be used for recently diverged populations of the same species
388 with small to moderate diversity. Furthermore, our method primarily targets
389 stochastic linkage, which is less important when comparing sequences from

390 differeent species (WEIGT *et al.* 2009; COCCO *et al.* 2018). Where both methods can be
391 potentially applied, such as calculating the fitness landscape of HIV Ab-binding
392 regions (LOUIE *et al.* 2018), our method is computationally faster, because it is local.
393 Indeed, we can consider one pair of loci at a time and not worry about simultaneous
394 optimization of $L^2/2$ parameters of the full interaction matrix. Also, it helps to avoid
395 the situation when the number of fitting parameter is too large, and the system is
396 over-defined.
397 To summarize, we proposed a technique to predict evolution of pairs of sites
398 and tease it out from stochastic linkage effects. We hope that our quasi-equilibrium
399 approach and further development of this technique will prove useful for all
400 researchers interested in finding fitness landscapes of various organisms from
401 genetic samples.
402
403 **METHODS**
404
405 **Wright-Fisher model for simulation.** We simulate the evolution of a haploid
406 population of N binary sequences (ACEVEDO *et al.*), where each genome site
407 (nucleotide position) numbered by i =1, 2, …, L has either K_i =0 or K_i =1, in discrete
408 generations (Wright-Fisher model). The processes included are random mutation at
409 rate μL, per genome, natural selection, random genetic drift, and an epistatic
410 network with a set strength and topology. Recombination is absent. Previous work
411 *in silico* shows that sufficient levels of recombination would enhance epistatic
412 detection (PEDRUZZI AND ROUZINE 2019). In the presence of epistasis, the average
413 progeny number (Darwinian fitness) of sequence [K_i] is set to e^W where
414
415 $$W = -\sum_{i=1}^{L} s_i K_i - \sum_{i<j}^{L} s_{ij} K_i K_j \quad (3)$$
416
417 $$s_{ij} = E_{ij}(s_i + s_j)T_{ij} \quad (4)$$
418
419 where the selection coefficients s_i and s_j represent the fitness costs of two

420 deleterious mutations that are partially compensated by each other. By definition, E_{ij}
421 is the degree of compensation of deleterious alleles at sites i and j. Values E = 0 and
422 1 represents no epistasis and full compensation, respectively.
423
424 **Quasi-equilibrium approximation**. In a broad parameter range, an adapting
425 asexual or partly sexual population represents a slowly-moving, narrow peak in
426 fitness space (ROUZINE *et al.* 2003; ROUZINE AND COFFIN 2005; ROUZINE *et al.* 2008;
427 ROUZINE AND COFFIN 2010; GOOD *et al.* 2012). Its speed is restricted by the formation
428 and fixation of rare beneficial mutations in the few best-fit genomes (the leading
429 edge). Because the fitness distribution moves slowly, the alleles entropy of the
430 mutation distribution over genomes has enough time to reach its current maximum,
431 restricted by the current average fitness of the population. This situation is called
432 "quasi-equilibrium". We verified the validity of Eq. 3 in a broad range of parameters
433 and initial conditions after time ~ $1/\langle s \rangle$ (PEDRUZZI *et al.* 2018). When the population
434 finally arrives in mutation-selection equilibrium, the presence of deleterious
435 mutation events compensates the natural selection and creates linkage equilibrium
436 and the mutation-selection balance. At this point, the method stops working.
437
438 **Linkage measure**. In our recent work (PEDRUZZI *et al.* 2018), we introduced a binary
439 measure of LD
440
441 $$\mathrm{UFE} = 1 - \frac{\log\left(f_{11}/f_{00}\right)}{\log\left(f_{01}f_{10}/{f_{00}}^{2}\right)} \quad (5)$$
442
443 where f_{00} , f_{01}, f_{10} f_{11} are the average-over-ensemble haplotype frequencies, where 0
444 corresponds to consensus sequence. As follows from Eqs. (1) and (2), UFE = E for an
445 isolated epistatic pair (PEDRUZZI *et al.* 2018).
446 UFE performs similarly to more widely used measures, such as Lewontin's D'
447 and Pearson correlation coefficient, r^2 (PEDRUZZI AND ROUZINE 2019) or mutual
448 information (WEIGT *et al.* 2009; COCCO *et al.* 2018). As compared to these measures,
449 UFE has the unique advantage of directly measuring the degree of mutual

compensation of two alleles E (Eq. 2), provided they are epistatically isolated. If they are a part of a network, this measure overestimates E. We calculate UFE for every potential epistatic link in our simulated sequence set (*Results*).

Tri-site test of false bonds. To test whether the interaction is direct, and not due to linkage or an indirect detour, we calculate for each suspected pair *i, j* the value

$$\mathrm{UFE}_{ij0} = 1 - \frac{\log\left(f_{110}/f_{000}\right)}{\log\left(f_{010}f_{100}/{f_{000}}^{2}\right)} \tag{6}$$

where the third 0 selects only for the sequences with the consensus allele 0 at a chosen neighbor site of the pair. We try all possible neighbor sites and calculate the minimum value of UFE_{ij0} over all neighbors. This method cuts the most important indirect correlation by "detour" for a site pair. It works for both linkage and a modest amount of indirect correlation.

To illustrate how this method works on indirect bonds, let us consider a simple analytic example in Fig. 1. We assume that some sites carry deleterious alleles with equal selection coefficient $s_i = s_0$ in Eq. 3, with a fixed epistatic strength $E_{ij} = E$. In this simple case, we can fully characterize a genome by the numbers of interacting allelic clusters of different size. To do so, let k_i define the number of clusters with i nodes and b_i bonds. Then, from Eq. 3, we can express fitness as a sum over clusters of different size (PEDRUZZI *et al.* 2018)

$$W \equiv -s_0 f_0 L = -s_0 \sum_{i=1}^{i_{max}} k_i (i - 2E b_i) \tag{7}$$

New notation f_0 represents the frequency of uncompensated mutations with total fitness W. The number of bonds b_i depends on the topology, and $b_1 = 0$, $b_2 = 1$. In our example with double arches, $b_3 = 2$ (Fig. 1).

As it has been shown previously in the literature on traveling wave (see *Introduction*), the population of genomes represents a slowly moving, narrow fitness peak. Because it in the state of quasi-equilibrium determined by its current

average fitness, at each moment of time, numbers k_i are determined by the condition that the entropy of the system is maximum given the value of fitness, Eq. 7. Entropy S is defined as the log number of sequence configurations

$$S = \log\left[\prod_{i=1}^{i_{max}} C_{L_i}^{k_i} (n_i)^{k_i}\right] \tag{8}$$

where L_i is the number of all possible locations for a cluster of size i, and n_i is the number of each cluster's configurations (shapes). The values of L_i and n_i depend on the network topology.

Previously, we applied this argument for several topologies to predict cluster numbers (PEDRUZZI *et al.* 2018). We showed, for the one in Fig. 1, that the numbers of clusters of size i = 1, 2 and 3 in, normalized as $f_i = k_i/L$, are related as

$$f_2 = \frac{1}{3} f_1^{2-2E}, \;\; f_3 = \frac{2}{3} f_1^{3-4E} \qquad E < \frac{3}{4} \tag{9}$$

The 1st and 3d site in each triplet in Fig. 1 do not interact directly, but only indirectly through site 2. For these two sites, the haplotype frequencies are

$$f_{11} = 3f_3 + f_1^2, \quad f_{10} = \frac{3}{2} f_2 + f_1 \tag{10}$$

When epistatic interaction is sufficiently strong, as given by the condition $E > 1/2$, triplets dominate over single alleles and their doubles, as given by $f_1 \ll f_2 \ll f_3$ (Eq. 9). In this case, we can approximate haplotype frequencies as

$$f_{11} \approx 2f_1^{3-4E}, \;\; f_{10} \approx \frac{1}{2} f_1^{2-2E}, \qquad \frac{1}{2} < E < \frac{3}{4} \tag{11}$$

Using covariance measure UFE_{ij} defined in Eq. 3, we obtain $\mathrm{UFE}_{ij} \approx \frac{2E-1}{2-2E}$. We observe that the indirect covariance between sites 1 and 3 is very strong, ~ 1. [In fact, covariance is approximately the same for directly interacting sites 1 and 2 (Fig.

1), for which we previously obtained $f_{11}^{dir} \approx f_1^{3-4E}, f_{10}^{dir} \approx \frac{1}{3} f_1^{2-2E}$, see (PEDRUZZI *et al.* 2018), *Supplement*, Eqs. (3.29) and (3.30)].

However, if we calculate UFE_{ij0} instead of UFE_{ij} by including only the sequences with majority allele "0" found at position 2, we obtain

$$f_{101} \approx f_1^2, \;\; f_{100} \approx f_1 \tag{10}$$

instead of Eq. 9, and Eq. 6 yields $\mathrm{UFE}_{ij0} = 0$. Thus, phantom covariance disappears, when we select only sequences with a majority allele between the sites of interest. This result is intuitively clear: by the definition of fitness (Eq. 1), only minority alleles "1" interact with each other, while majority alleles are neutral background. The same method turns out to be extremely effective for eliminating false bonds created by linkage (see *Results*, Fig. 3).

Sequence preparation. We have applied the three-way test to influenza virus. Evolution of the virus in population has been mapped recently to traveling-wave models (ROUZINE AND ROZHNOVA 2018; YAN *et al.* 2019), which justifies the use of quasi-equilibrium approximation dependent on current average fitness (see Intro for the explanation). We performed a multiple progressive alignment for amino acid sequences obtained for two proteins of Influenza A virus strain H1N1, Neuraminidase (NA) and Hemagglutinin (HA) from public database https://www.fludb.org.

We have used all the sequences discovered in the database for the period 2000-2010, which all come from various geographic locations. Our aim was to understand the transition between the protein variants before and after pandemic of 2009, which have 80% of mutual homology. We aimed at discovering only the epistatic sub-network related to that transition, and are not interested in any other epistatic interactions in these proteins. For this end, we compared worldwide samples of sequences from the two strains. We randomly sampled similar amounts of sequences from the first and second strains, and re-sampled them several

539 hundred times (*Results*). We also checked the robustness of the results to exact
540 sampling size (see *Results*).
541 Pairwise distances between sequences were computed using pairwise
542 alignment with the Gonnet scoring matrix implemented in MATLAB. To calculate the
543 guide tree we used the neighbor-joining method assuming equal variance and
544 independence of evolutionary distance estimates. The obtained consensus served as
545 a universal reference to binarize data sequences. Before applying the detection
546 algorithm, the protein sequences were binarized, by direct comparison of each
547 sequence to the consensus. Each amino-acid residue was set to 0 or 1 for consensus
548 or non–consensus. Pooling all mutations per site ignores the chemistry of
549 substitutions. In exchange, this approach greatly reduces the number of haplotype
550 combinations and increases the sensitivity by effectively increasing the haplotype
551 frequencies. Next, we measure the mutational frequency for each sequence, for each
552 site. The subset of low-diversity sequences with allelic frequency below a cut-off d_v
553 is randomly sampled and down-weighted according to the coefficient D_w. Then, we
554 determined the average pairwise and triple haplotype frequencies for pairs and
555 triplets of sites, to be used in the method described above.
556
557 **3D structures**. To characterize the three-dimensional network of epistatic
558 interaction, we used software package ChimeraX from internet site
559 https://www.rbvi.ucsf.edu/chimerax/.
560
561 **Data availability.** Influenza sequence data are from public database
562 https://www.fludb.org.
563
564 **Acknowledgements.** We thank Martin Weigt and Alessandra Carbone for useful
565 comments. This research has been funded by Agence Nationale de la Recherche
566 grant J16R389 to IMR, http://www.agence-nationale-recherche.fr/.

567
568 **References**

Acevedo, A., L. Brodsky and R. Andino, 2014 Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature 505: 686-690.

Barlukova, A., G. Pedruzzi and I. M. Rouzine, 2020 The evolutionary origin of the universal distribution of mutation fitness effect https://biorxiv.org/cgi/content/short/867390v1.

Barton, N. H., 1995 Linkage and the limits to natural selection. Genetics 140: 821-841.

Brunet, E., I. M. Rouzine and C. O. Wilke, 2008 The stochastic edge in adaptive evolution. Genetics 179: 603-620.

Byarugaba, D. K., B. Erima, M. Millard, H. Kibuuka, L. Lkwago *et al.*, 2016 Whole-genome analysis of influenza A(H1N1)pdm09 viruses isolated in Uganda from 2009 to 2011. Influenza Other Respir Viruses 10: 486-492.

Carlborg, O., L. Jacobsson, P. Ahgren, P. Siegel and L. Andersson, 2006 Epistasis and the release of genetic variation during long-term selection. Nat Genet 38: 418-420.

Chen, C. C., H. Schwender, J. Keith, R. Nunkesser, K. Mengersen *et al.*, 2011 Methods for identifying SNP interactions: a review on variations of Logic Regression, Random Forest and Bayesian logistic regression. IEEE/ACM Trans Comput Biol Bioinform 8: 1580-1591.

Cocco, S., C. Feinauer, M. Figliuzzi, R. Monasson and M. Weigt, 2018 Inverse statistical physics of protein sequences: a key issues review. Rep Prog Phys 81: 032601.

Cordell, H. J., 2009 Detecting gene-gene interactions that underlie human diseases. Nature reviews. Genetics 10: 392-404.

Desai, M. M., and D. S. Fisher, 2007 Beneficial mutation selection balance and the effect of linkage on positive selection. Genetics 176: 1759-1798.

Felsenstein, J., 1974 The evolutionary advantage of recombination. Genetics 78: 737-756.

Fisher, R. A., 1930 *The genetical theory of natural selection*. Clarendon Press, 1958, Oxford, United Kingdom.

Gonzalez-Ortega, E., E. Ballana, R. Badia, B. Clotet and J. A. Este, 2011 Compensatory mutations rescue the virus replicative capacity of VIRIP-resistant HIV-1. Antiviral Res 92: 479-483.

Good, B. H., I. M. Rouzine, D. J. Balick, O. Hallatschek and M. M. Desai, 2012 Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. Proc Natl Acad Sci U S A 109: 4950-4955.

Hallatschek, O., 2011 The noisy edge of traveling waves. Proc Natl Acad Sci U S A 108: 1783-1787.

Handel, A., R. R. Regoes and R. Antia, 2006 The role of compensatory mutations in the emergence of drug resistance. PLoS Comput Biol 2: e137.

Hill, W. G., and A. Robertson, 1966a The effect of linkage on limits to artificial selection. Genet. Res. 8: 269-294.

Hill, W. G., and A. Robertson, 1966b The effect of linkage on limits to artificial selection. Genet Res 8: 269-294.

614 Kimura, M., 1994 *Population genetics, molecular evolution, and the neutral theory.*
615 *Selected papers.* The University of Chicago Press, Chicago.
616 Kryazhimskiy, S., J. Dushoff, G. A. Bazykin and J. B. Plotkin, 2011 Prevalence of
617 epistasis in the evolution of influenza A surface proteins. PLoS Genet 7**:**
618 e1001301.
619 Levin, B. R., V. Perrot and N. Walker, 2000 Compensatory mutations, antibiotic
620 resistance and the population genetics of adaptive evolution in bacteria.
621 Genetics 154**:** 985-997.
622 Louie, R. H. Y., K. J. Kaczorowski, J. P. Barton, A. K. Chakraborty and M. R. McKay,
623 2018 Fitness landscape of the human immunodeficiency virus envelope
624 protein that is targeted by antibodies. Proc Natl Acad Sci U S A 115**:** E564-
625 E573.
626 Muller, H. J., 1932 Some genetic aspects of sex. Am. Nat. 66**:** 118-128.
627 Neher, R. A., and O. Hallatschek, 2013 Genealogies of rapidly adapting populations.
628 Proc Natl Acad Sci U S A 110**:** 437-442.
629 Neher, R. A., B. I. Shraiman and D. S. Fisher, 2010 Rate of adaptation in large sexual
630 populations. Genetics 184**:** 467-481.
631 Neverov, A. D., S. Kryazhimskiy, J. B. Plotkin and G. A. Bazykin, 2015 Coordinated
632 Evolution of Influenza A Surface Proteins. PLoS Genet 11**:** e1005404.
633 Nijhuis, M., R. Schuurman, D. de Jong, J. Erickson, E. Gustchina *et al.*, 1999 Increased
634 fitness of drug resistant HIV-1 protease as a result of acquisition of
635 compensatory mutations during suboptimal therapy. AIDS 13**:** 2349-2359.
636 Otte, A., A. C. Marriott, C. Dreier, B. Dove, K. Mooren *et al.*, 2016 Evolution of 2009
637 H1N1 influenza viruses during the pandemic correlates with increased viral
638 pathogenicity and transmissibility in the ferret model. Sci Rep 6**:** 28583.
639 Pedruzzi, G., A. Barlukova and I. M. Rouzine, 2018 Evolutionary footprint of
640 epistasis. PLoS Comput Biol 14**:** e1006426.
641 Pedruzzi, G., and I. M. Rouzine, 2019 Epistasis detectably alters correlations
642 between genomic sites in a narrow parameter window. PLoS ONE in press.
643 Piana, S., P. Carloni and U. Rothlisberger, 2002 Drug resistance in HIV-1 protease:
644 Flexibility-assisted mechanism of compensatory mutations. Protein Sci 11**:**
645 2393-2402.
646 Rice, S. H., 2004 *Evolutionary Theory: Mathematical And Conceptual Foundations*.
647 Sinauer Associated.
648 Rouzine, I. M., E. Brunet and C. O. Wilke, 2008 The traveling-wave approach to
649 asexual evolution: Muller's ratchet and speed of adaptation. Theor Popul Biol
650 73**:** 24-46.
651 Rouzine, I. M., and J. M. Coffin, 1999 Search for the mechanism of genetic variation in
652 the pro gene of human immunodeficiency virus. J Virol 73**:** 8167-8178.
653 Rouzine, I. M., and J. M. Coffin, 2005 Evolution of human immunodeficiency virus
654 under selection and weak recombination. Genetics 170**:** 7-18.
655 Rouzine, I. M., and J. M. Coffin, 2010 Multi-site adaptation in the presence of
656 infrequent recombination. Theor Popul Biol 77**:** 189-204.
657 Rouzine, I. M., and G. Rozhnova, 2018 Antigenic evolution of viruses in host
658 populations. PLoS Pathog 14**:** e1007291.

659 Rouzine, I. M., J. Wakeley and J. M. Coffin, 2003 The solitary wave of asexual
660 evolution. Proc Natl Acad Sci U S A 100**:** 587-592.
661 Seibert, C. W., S. Rahmat, F. Krammer, P. Palese and N. M. Bouvier, 2012 Efficient
662 transmission of pandemic H1N1 influenza viruses with high-level oseltamivir
663 resistance. J Virol 86**:** 5386-5389.
664 Takahashi, T., J. Song, T. Suzuki and Y. Kawaoka, 2013 Mutations in NA that induced
665 low pH-stability and enhanced the replication of pandemic (H1N1) 2009
666 influenza A virus at an early stage of the pandemic. PLoS One 8**:** e64439.
667 Ueki, M., and H. J. Cordell, 2012 Improved statistics for genome-wide interaction
668 analysis. PLoS Genet 8**:** e1002625.
669 Walczak, A. M., L. E. Nicolaisen, J. B. Plotkin and M. M. Desai, 2012 The structure of
670 genealogies in the presence of purifying selection: a fitness-class coalescent.
671 Genetics 190**:** 753-779.
672 Wei, W.-H., G. Hemani and C. S. Haley, 2014 Detecting epistasis in human complex
673 traits. Nature Reviews Genetics 15**:** 722.
674 Weigt, M., R. A. White, H. Szurmant, J. A. Hoch and T. Hwa, 2009 Identification of
675 direct residue contacts in protein-protein interaction by message passing.
676 Proc Natl Acad Sci U S A 106**:** 67-72.
677 Weissman, D. B., M. M. Desai, D. S. Fisher and M. W. Feldman, 2009 The rate at which
678 asexual populations cross fitness valleys. Theor Popul Biol 75**:** 286-300.
679 Wu, N. C., A. J. Thompson, J. Xie, C. W. Lin, C. M. Nycholat *et al.*, 2018 A complex
680 epistatic network limits the mutational reversibility in the influenza
681 hemagglutinin receptor-binding site. Nat Commun 9**:** 1264.
682 Yan, L., R. A. Neher and B. I. Shraiman, 2019 Phylodynamic theory of persistence,
683 extinction and speciation of rapidly adapting pathogens. Elife 8.
684 Zuk, O., E. Hechter, S. R. Sunyaev and E. S. Lander, 2012 The mystery of missing
685 heritability: Genetic interactions create phantom heritability. Proc Natl Acad
686 Sci U S A 109**:** 1193-1198.
687

688

FIGURES AND CAPTIONS

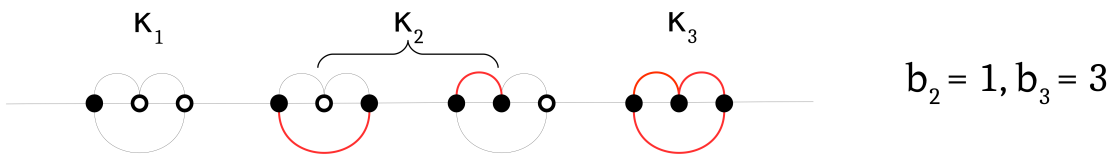

Fig. 1. A simplest epistatic network illustrating the effect of indirect interactions. Black and white circles are deleterious and favorable alleles, respectively, arches show direct (potential) epistatic interactions, red arches are actual interactions between deleterious alleles. Numbers k_i denote the numbers of *i*-clusters in a genome.

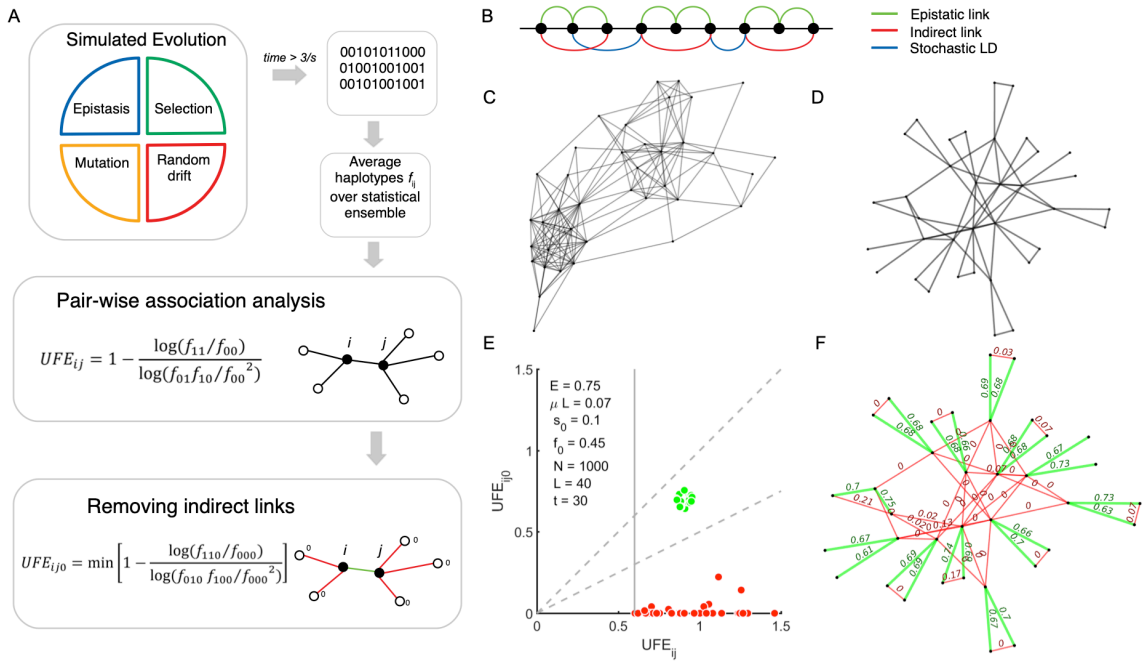

Fig. 2 Schematic diagram of the method and its training on simulated sequence data. A. The computer model of asexual evolution includes the factors of random mutation, selection, epistasis, and random genetic drift. Pairwise haplotype frequencies f_{ij} are averaged over an ensemble of independent simulation runs (populations). The strength of interactions UFE_{ij} is calculated. The indirect links and the residual stochastic LD are excluded by using triple-site haplotype frequencies, UFE_{ij0}. **B.** Epistatic links (green), indirect links (red), and stochastic LD (blue). **C-D.** The network of strong candidate epistatic interactions predicted (**C**) from a single population and (**D**) after averaging over 200 populations. **E.** Scatter plot of 3-locus haplotype min(UFE_{ij0}) shown against UFE_{ij} for the pairs identified in (**D**). Dashed sector (green): Predicted strong, direct interactions. **F.** Predicted network accurately recapitulates the pre-set epistatic network. Parameters are shown (**E**).

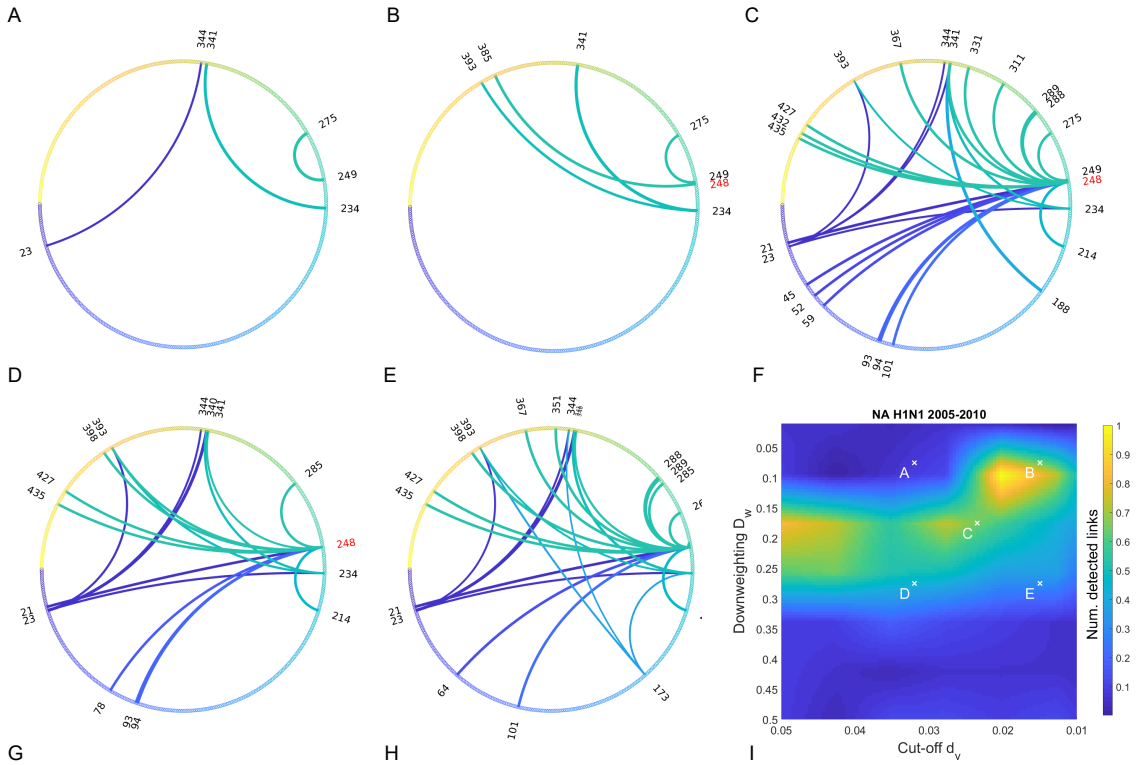

Fig. 3 Epistatic network predicted from sequence data on surface protein sequences of Influenza A H1N1 obtained between year 2005 and year 2010. The circular diagrams show the network of interaction for (**A-E**) neuraminidase and (**G-K**) hemagglutinin. (**F, L**) Sequences with homology to the consensus less than d_v are randomly re-sampled many times, with their number downweighted by coefficient D_w. (**F, L**) 2D heatmaps showing the total number of links as a function of d_v (X-axis) and D_w (Y-axis). Different versions of wheels in **A-E** (or **G-K**) correspond to different choices of D_w and d_v shown by crosses in **F, L.**

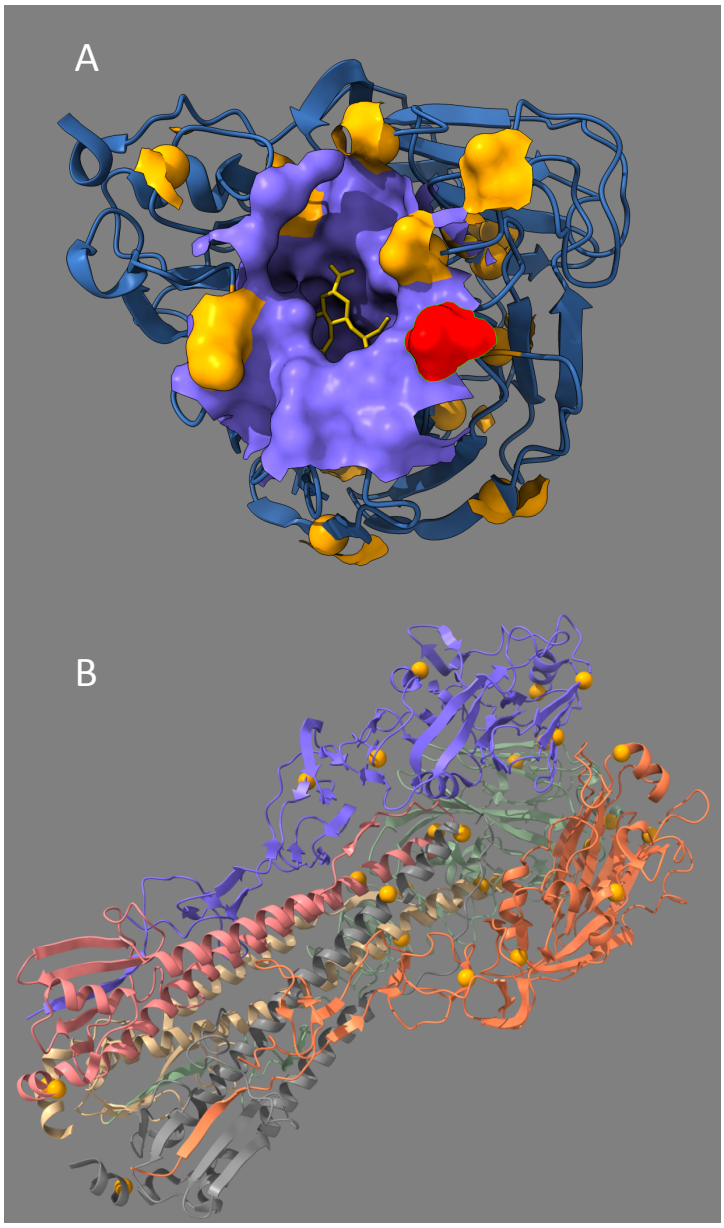

Fig. 4 Structural location of the predicted epistasis network for the two influenza virus envelope proteins. The figure shows the three-dimensional structure of Influenza A H1N1 proteins (**A**) neuraminidase (PDB ID code 4QVZ) and (**B**) hemagglutinin (PDB ID code 1RUZ). Colored spheres represent predicted epistatic residues from Fig. 3D for (**A**) and Fig. 3J for (**B**). Red sphere: Predicted primary mutation (residue 248 in Fig 3D). Orange spheres: Compensatory residues from Fig. 3D and 3J.