Statistical significance of spectral identifications of non-linear peptides

Hosein Mohimani,[†] Sangtae Kim,[‡] and Pavel A. Pevzner[*,‡]

Department of Electrical and Computer Engineering, UC San Diego, and Department of Computer Science and Engineering, UC San Diego

E-mail: ppevzner@ucsd.edu

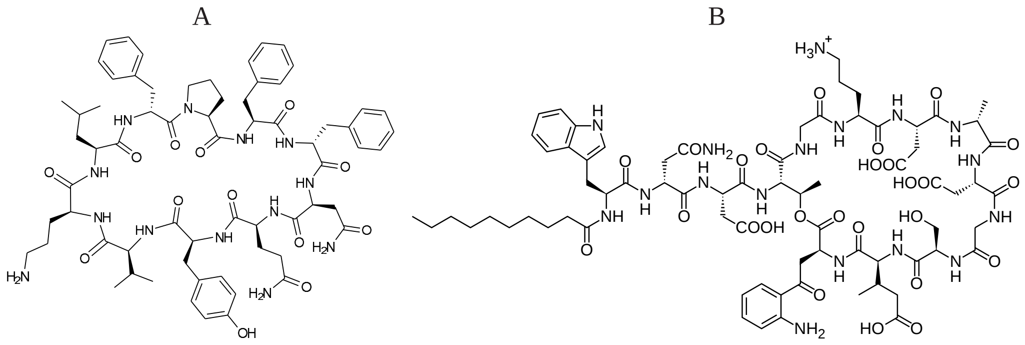

Fig. S1: (A) Structure of Tyrocidine. (B) Structure of Daptomycin.

Appendix

While the DPR algorithm is guaranteed to converge to the equilibrium distribution, we have no estimate of the convergence rate, i.e., how many iterations does it take for the markov chain to reach equilibrium distribution. Convergence rate of the markov chain is critically dependent on the choice of $RandomTransition(Peptide)$, and with a bad choice of $RandomTransition(Peptide)$,

[*] To whom correspondence should be addressed

[†] Department of Electrical and Computer Engineering, UC San Diego

[‡] Department of Computer Science and Engineering, UC San Diego

convergence can be so slow that there would be no improvement for DPR as compared to naive Monte Carlo simulations. We explain this concept with the following example. Consider the two score transition probability matrices P and Q shown in **Table 4** (p is a small value). While both transition matrices have similar equilibrium distribution $\pi = (1/(1+2p+4p^2), 2p/(1+2p+4p^2), 4p^2/(1+2p+4p^2))$, they are different in the sense that for the former, there is a path $s_1 \to s_2 \to s_3$ going from the most common state to the rarest state, where the probability of each transition is proportional to p, while in the latter case no such path exist, and the probability of going to the most rare state is always quadratic with p, no matter which state the transition is originated from. By choosing oversampling factors $\mu = (1, 1/2p, 1/4p^2)$, the modified transition probability matrices calculated from DPR paper are shown in **Table 4**.

Table S1: The score transition probability matrices P and Q, and modified probability transition matrices P' and Q', coming from performing DPR on P and Q.

$$\mathbf{P} = \begin{bmatrix} 1-p & p & 0 \\ 0.5 & 0.5-p & p \\ 0 & 0.5 & 0.5 \end{bmatrix} \quad \mathbf{Q} = \begin{bmatrix} 1/(1+2p+4p^2) & 2p/(1+2p+4p^2) & 4p^2/(1+2p+4p^2) \\ 1/(1+2p+4p^2) & 2p/(1+2p+4p^2) & 4p^2/(1+2p+4p^2) \\ 1/(1+2p+4p^2) & 2p/(1+2p+4p^2) & 4p^2/(1+2p+4p^2) \end{bmatrix}$$

$$\mathbf{P'} = \begin{bmatrix} 1-p & p & 0 \\ p & 1-2p & p \\ 0 & p & 1-p \end{bmatrix} \quad \mathbf{Q'} = \begin{bmatrix} 1/(1+2p+4p^2) & 2p/(1+2p+4p^2) & 4p^2/(1+2p+4p^2) \\ 2p/(1+2p+4p^2) & 1/(1+2p+4p^2) & 4p^2/(1+2p+4p^2) \\ 4p^2/(1+2p+4p^2) & 4p^2/(1+2p+4p^2) & (1+2p-4p^2)/(1+2p+4p^2) \end{bmatrix}$$

The eigenvalues of P' are $1, 1-p, 1-3p$, while eigenvalues of Q' are $1, 1-12p^2/(1+2p+4p^2), 1-2p/(1+2p+4p^2)$. The convegence rate of each markov chain is determined by the largest non-unity eigenvalue of their matrices, which means in the former case equilibrium distribution can be reached in the number of samples growing by $1/p$, while in the latter case it grows by $1/p^2$. Finally, note that the number of random samples that a crude monte carlo approach would require for estimating such a probability distribution is proportional to $1/p^2$. This means while using DPR with *RandomTransition* that gives score transition probability matrix Q has no overall payback as compared to naive Monte Carlo, transition probability matrix P can greatly reduce the number of samples required for accurately estimate the probability distribution.

In general an effective *RandomTransition* should have the following two properties to be effective. First, it should make the whole space of all peptides connected. Second, its score transition probability matrix should have paths from most common states to rarest states, where each edge has a significant probability (e.g., larger than 10^{-6}). Then, it would be possible to estimate the equilibrium probability distribution of such a matrix in just several million iterations.