

Abstract

Amino acid substitution matrices are widely available to the biological community. They specify the chance of one amino acid being substituted for another via mutation on DNA (or RNA). These substitution matrices can be interpreted as a directed graph of mutations. Although there is some physico-chemical requirement for particular codons to specify particular amino acids, one could infer that the substitution matrix that we observe has some evolutionary advantage: beneficial mutations from one amino acid to another are more likely to occur.

Given this reasoning, a core component of any evolving string-based artificial chemistry (for example typogenetics, avida) should be a substitution matrix which is used to determine the mutation sequence between tokens. This concept is under-represented in previous work on string-based chemistries. In the work reported here, we show the network properties that characterise an amino acid substitution matrix, and discuss how to apply them to artificially generated mutation sequences.

We calculate robust network statistics for two classes of substitution matrices. The first class of matrix concern amino acids in the living cell. The second class is a set of matrices with a similar number of nodes and edges to the biological class, but with randomised connectivity. Comparison of these two classes shows that the biological substitution matrices have significant structure, which corresponds with the functional properties (defined here as the contribution of the individual amino acid to the structure and function of the protein) of the amino acids in the network.

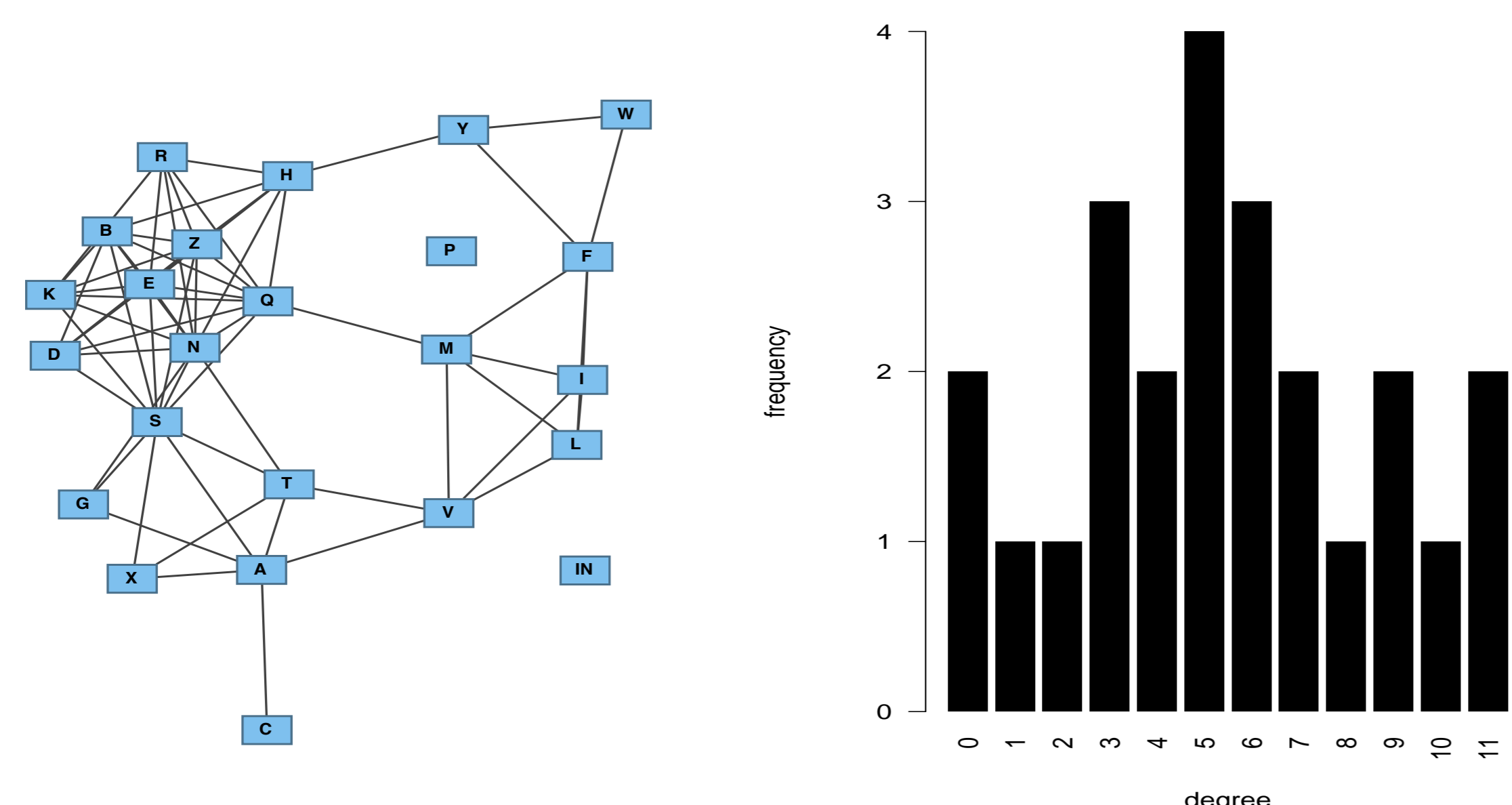
In earlier work, we described a simple heuristic for creating a substitution matrix for an artificial chemistry. We compare the network statistics of this matrix with the biological and random networks described above. Having established that biological substitution matrices record a beneficial mutation trajectory, we discuss how to construct substitution matrices which confer beneficial mutation trajectories in evolvable artificial chemistries.

Amino Acid Substitution Networks

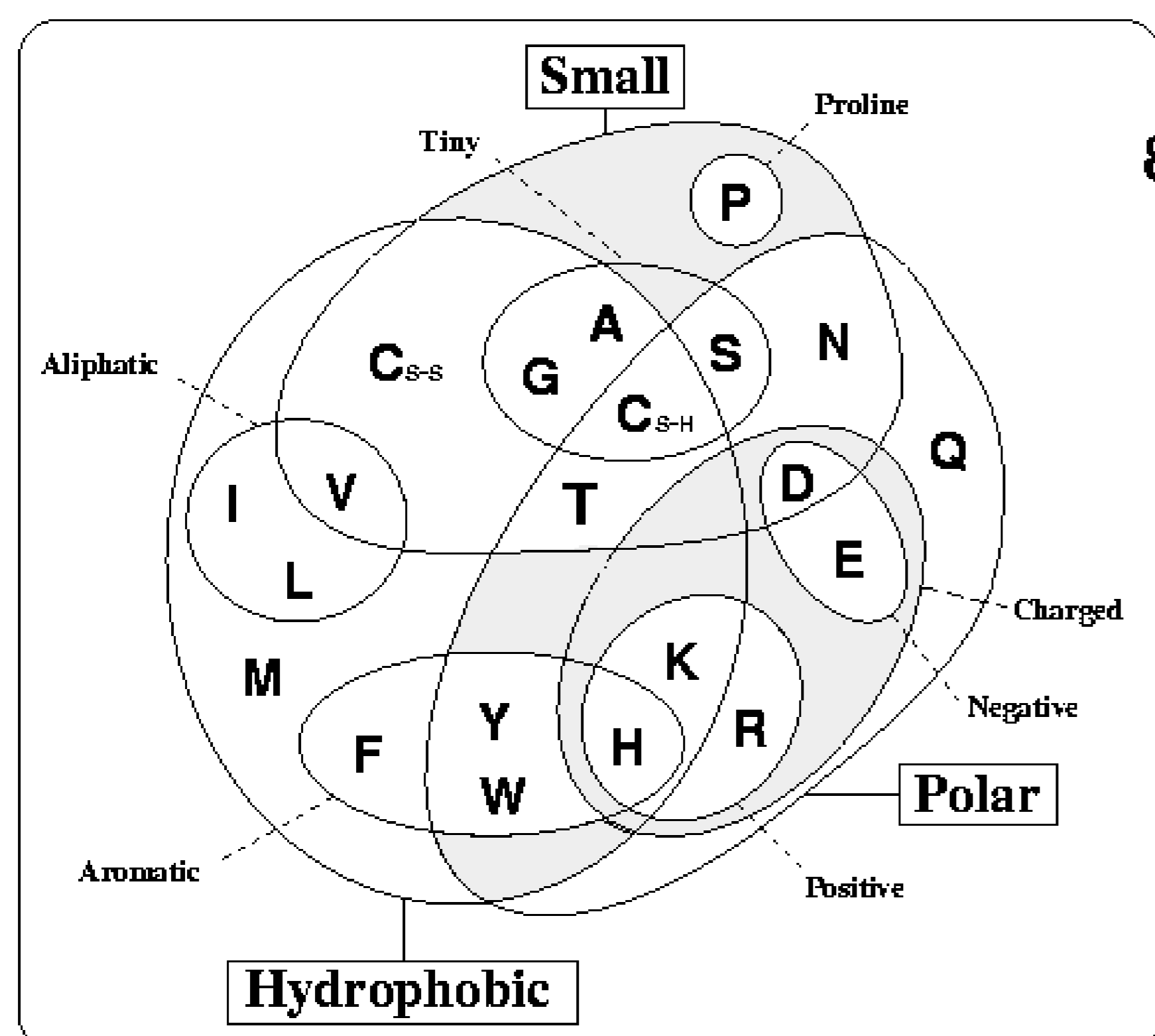
Amino acid substitution matrices assign a score to mismatches between sequences of amino acids. These scores are used to measure the mutation "distance" between proteins and have a variety of applications in biology. The Blosum62 substitution matrix [Henikoff and Henikoff, 1992] is a popular measure of this similarity and takes the form:

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	V	B	Z	X	Y	I	N		
A	4	-1	-2	-2	0	-1	0	-2	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	4					
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4			
N	-2	0	6	1	-3	0	0	0	1	-3	0	-2	-3	-2	1	0	-4	-2	-3	0	-1	-4					
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4			
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-2	-2	-1	-3	-2	-4					
Q	-1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4				
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	1	4	-1	-4				
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4			
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-1	-2	-2	2	3	0	-1	-4		
I	-1	-3	-3	-3	-1	-3	-4	-3	4	2	3	4	0	-3	-2	-1	-3	-1	-3	-3	-3	-1	-4				
L	-1	-2	-3	-4	-1	-2	-3	-4	2	4	2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4				
K	-1	2	0	-1	-3	1	1	-2	-1	-3	2	5	-1	-3	1	0	-1	-3	-2	0	1	-1	-4				
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	-3	-1	-1	-4				
F	-2	-3	-3	-2	-3	-3	-1	0	0	3	0	6	-2	-2	1	3	-1	-3	-3	-1	-4						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	4	7	-1	-4	-3	-2	-2	-1	-2	-4				
S	1	-1	1	0	-1	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	0	0	0	-4					
T	0	-1	0	-1	-1	-1	-2	-2	-1	-1	-1	-1	1	5	-2	0	-1	-1	0	-1	0	-4					
W	-3	-3	-4	-1	-2	-3	-2	-3	-2	-3	-3	-1	1	-4	-3	11	2	-3	-4	-3	-2	-4					
Y	-2	-2	-2	-3	-2	-1	-2	-3	-2	-1	-2	1	3	-2	-2	2	7	-1	-3	-2	-1	-4					
V	0	-3	-3	-3	-1	-2	-3	-3	3	1	-2	1	-2	-2	0	-3	-1	4	-3	-2	-1	-4					
B	-2	-1	3	4	-3	0	-1	-1	0	-3	4	0	-3	-2	0	-1	-4	-3	4	1	-1	-4					
Z	-1	0	0	-1	-3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-1	4	-1	-4					
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4					
I	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4					

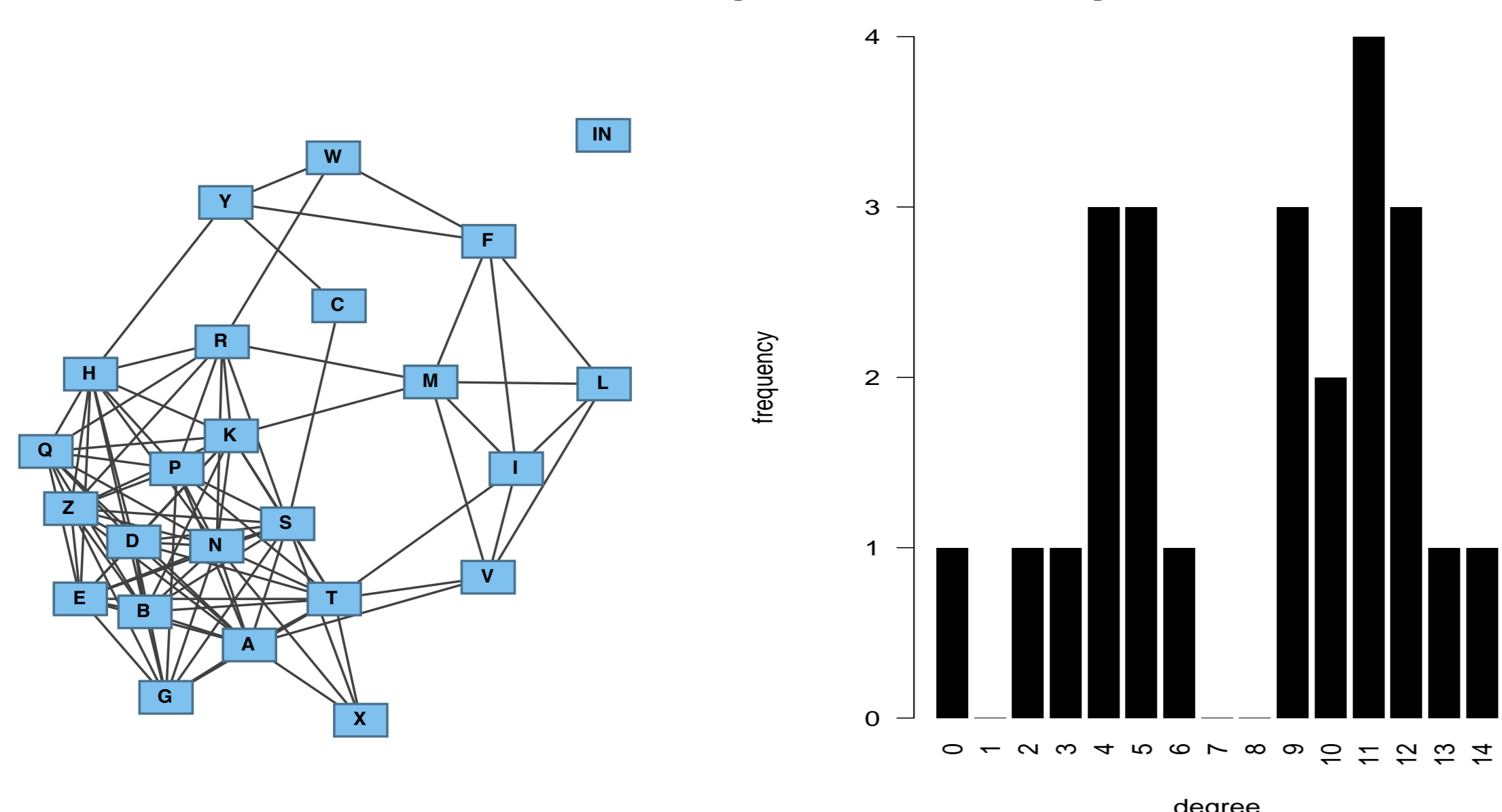
We can interpret this matrix as the specification of a network of mutational pathways by setting a threshold on the substitution scores. Any entry scoring above zero constitutes a likely mutation between amino acid nodes, and forms an edge between them. Below, we show the network that can be constructed for the Blosum62 matrix. We can measure statistics of these matrices to allow us to compare them to other matrices. The degree of a node is the number of connections it has to another node. The degree distribution is the probability distribution of these degrees, shown on the right below.



The network above was calculated solely from the Blosum substitution matrix, yet it bears striking resemblance to the diagram of amino acid properties from [Betts and Russell, 2003]. The edges in the network above form a graph whose shape approximates the functional groupings below:

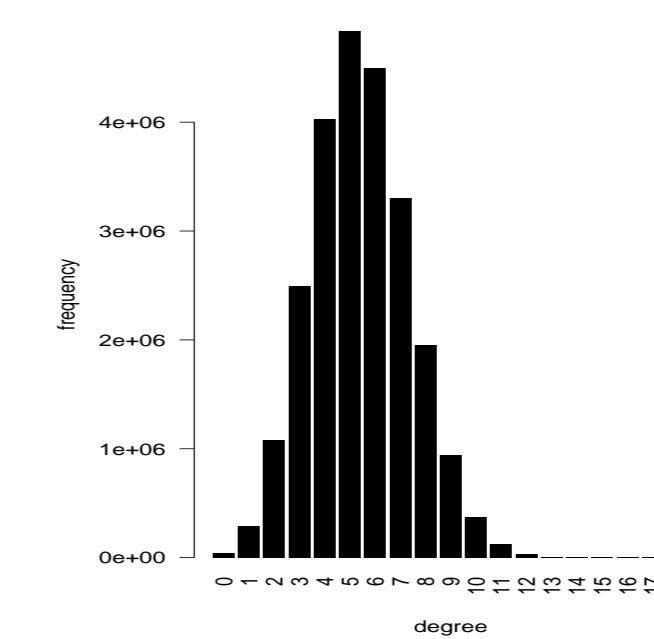


It is also similar to the PAM250 substitution matrix [Dayhoff et al., 1978]:

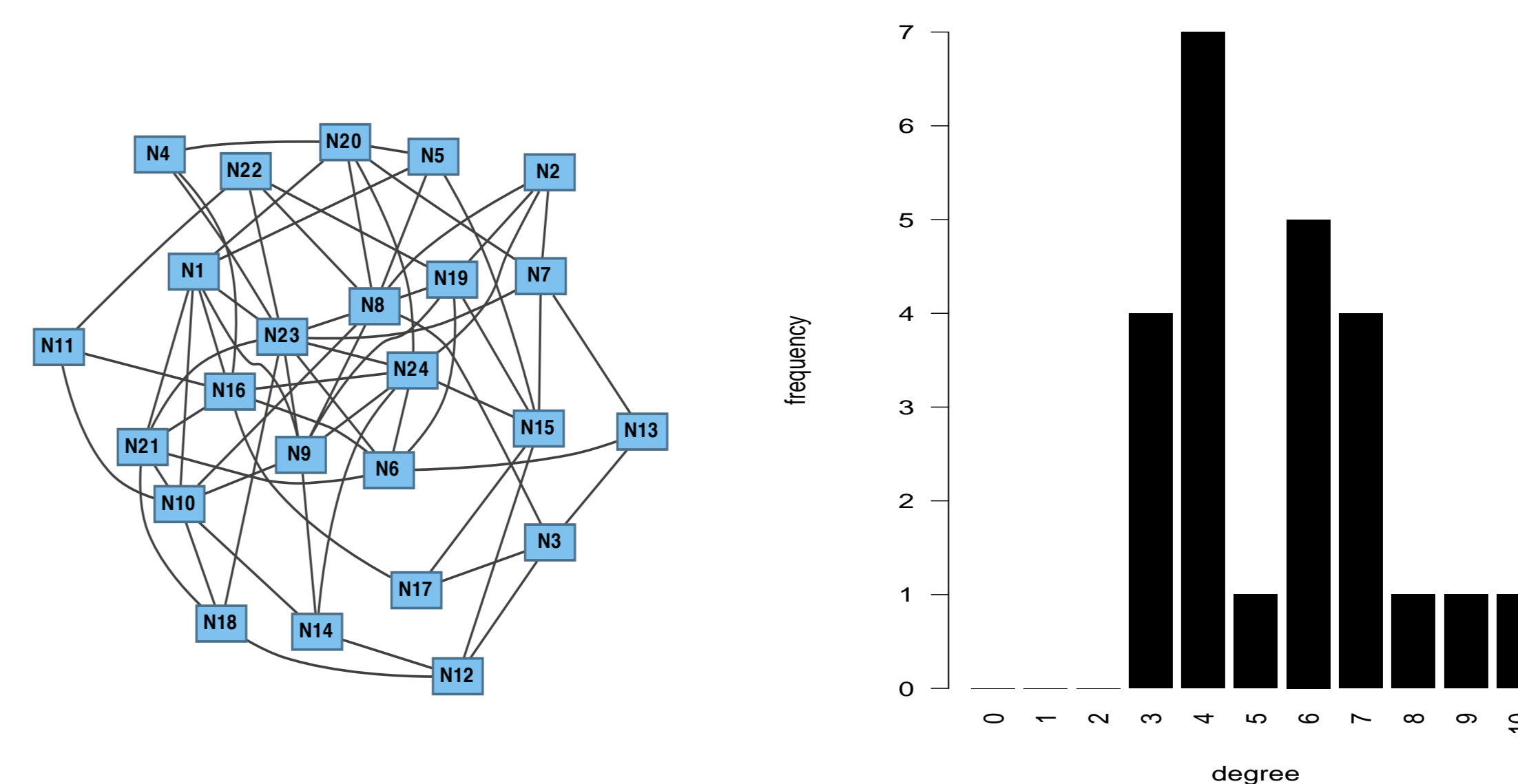


Random Matrices

To identify whether there was functional structure in the substitution networks, we constructed random networks with equal numbers of nodes and edges to the Blosum62 network described to the left. A compilation of degree distributions for 1,000,000 randomly connected networks is shown below:



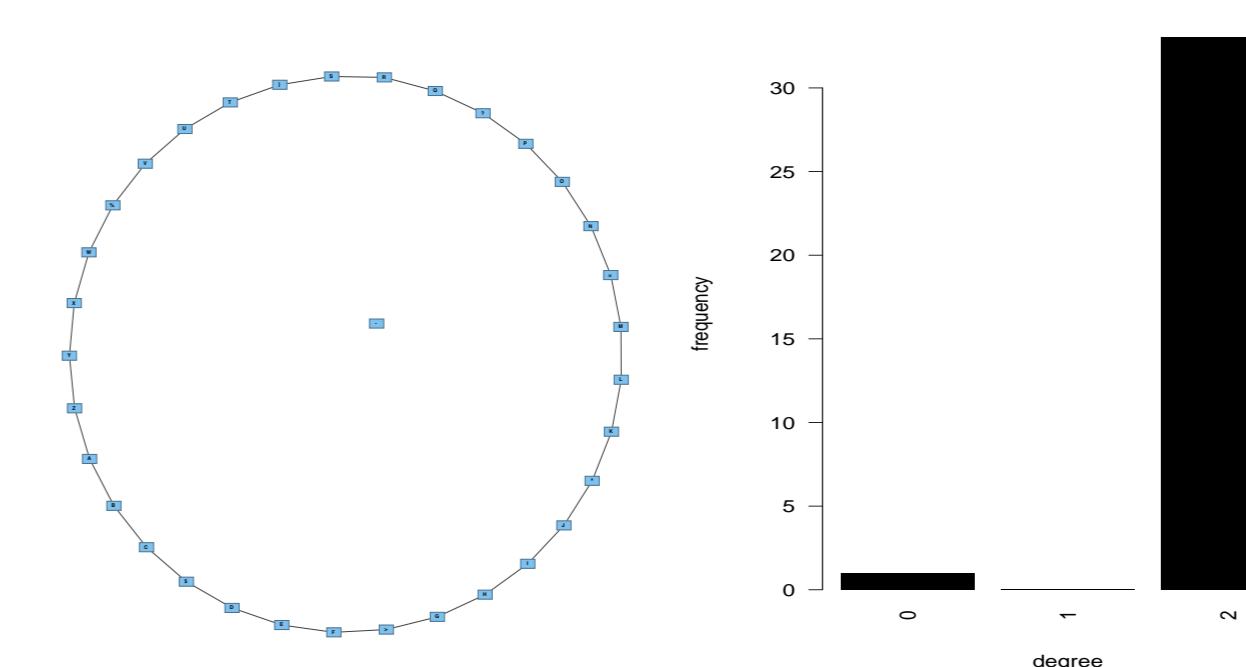
These networks we constructed using the same entries in the lower triangle of the Blosum62 substitution matrix, but allocated randomly. The figure illustrates that the degree distribution for randomly-connected matrices follow a binomial distribution. An example network and its distribution is shown below:



Visual comparison indicates that the amino acid substitution matrices have non-random connection patterns, that closely link related functionality between different nodes. Note that a measure of assortativity of these networks could reveal the structure in the amino acid networks more clearly.

Substitution matrices for Artificial Chemistries

In artificial chemistries such as Typogenetics, Avida and Tierra, mutation from one token to another is essentially random. Yet all of these chemistries make use of instruction sets whose tokens have related functionality. Ray [Ray, 1994] investigated several instruction sets and showed a variety of different categories (memory movement, calculation, sensory) of instruction, but with essentially random mutation pathways between them. The substitution network would thus follow a random distribution as shown above. In the Stringmol chemistry [Hickinbotham et al., 2010], mutations flow through a circular network, with the following degree distribution:



We argue that both of these mutation schemes are sub-optimal with regards to the substitution patterns between chemical tokens. The artificial chemical systems should implement mutation schemes which emphasise (but do not guarantee) substitutions between tokens of related function. This would change the dynamics of mutation, and encode a more secure exploration of the fitness landscape. Although automating the construction of functional substitution matrices is a difficult challenge, it should be relatively straightforward to encode substitution between related tokens by hand. Although this strategy is still likely to be sub-optimal, it is also likely to be significantly better than a strategy of random mutations, as illustrated by the biological analogue.

References

- 1. Betts, M. J. and Russell, R. B. (2003). Amino acid properties and consequences of substitutions. *Bioinformatics for Geneticists*, 317.
- 2. Dayhoff, M., Schwartz, R., and Orcutt, B. (1978). A model of evolutionary change in proteins. *Atlas of protein sequence and structure*, 5:345-358.
- 3. Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22):10915-10919.
- 4. Hickinbotham, S., Clark, E., Stepney, S., Clarke, T., Nellis, A., Pay, M., and Young, P. (2010). Specification of the stringmol chemical programming language version 0.2. Technical Report YCS-2010-458, Univ. of York.
- 5. Ray, T. S. (1994). Evolution, complexity, entropy and artificial reality. *Phys. D*, 75(1-3):239-263.