Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Infectious disease informatics

DOI: 10.1533/9781908818416.99

Abstract: Throughout history, infectious diseases have posed a serious burden to mankind. More recently, there has been an alarming increase in drug-resistant microbes. Furthermore, new pathogens are emerging due to microbial evolution and adaptation. The spread of these diseases is a result of pathogen mutations and changes in human behavior patterns. Then, there are diseases that are lurking in the background, waiting for the right conditions before they strike again. In the war against these diseases, we have come to understand the behaviors of microbes in a heterogeneous world and the mechanisms governing disease transmission. These works have profoundly shaped modern knowledge of emerging and re-emerging infections. More recently, computational techniques have led the way into this new era by allowing rapid high-throughput analysis of pathogens which was previously not possible using traditional laboratory techniques. This chapter introduces methods in mathematical modeling, computational biology, and bioinformatics that have been used to study infectious diseases.

Key words: infectious disease modeling, infectious disease informatics, sequence analysis, information entropy, molecular evolution.
7.1 Infectious diseases in history

Epidemics, pandemics, and outbreaks of infectious diseases are regular features of life on earth. In 430 BC, Thucydides described the very first pandemic in recorded history – the Athenian plague that reportedly killed up to one-half of the citizens of Athens. In AD 541–2, an outbreak occurred in the Byzantine Empire, causing 10,000 deaths every day. The outbreak, named the Justinian plague after the reigning emperor Justinian I, resulted in over 100 million deaths and wiped out nearly half the inhabitants of the city. In 1348–50, the plague returned to Europe under the name of the Black Death, killing up to 60% of the continent’s population. In March 1918, an influenza outbreak was first reported in a US military camp in Kansas. The outbreak, later known as the “Spanish flu,” subsequently spread and infected up to a billion people, or half the world’s population at the time, causing some 50 million deaths within six months. All over the world, changes in socio-economic, demographic and environmental factors brought about by urbanization and industrialization have led to the resurgence of old and new infectious diseases.

Over the past 40 years, there has been an alarming increase in drug-resistant microbes in diseases such as malaria and tuberculosis. Furthermore, the world is also witnessing the emergence of more new pathogens due to microbial evolution and adaptation. These include the Marburg virus in 1967, Lassa fever virus in 1969, rotavirus in 1973, Ebola virus in 1976, human immunodeficiency virus (HIV) in 1981, hepatitis C virus (HCV) in 1989, hantavirus in 1993, and the severe acute respiratory syndrome coronavirus (SARS-CoV) in 2002. The spread of these diseases is a result of pathogen mutations and changes in human behavior patterns, including lifestyles, environmental degradation, travel, and drug use.
Then, there are diseases that are lurking in the background, waiting for the right conditions before they strike again. In 1999, West Nile virus re-emerged in New York and spread across the United States to Long Island, Connecticut, Maryland, Florida, California, Arizona, and Colorado, with over 4100 reported cases and 280 associated deaths within a span of five years. Previously known to be a mild disease, the re-emergence of epidemic Chikungunya virus (CHIKV) in Africa, Indian Ocean, South-East Asia, Pacific, North America, and Europe in the past decade has caused severe morbidity with some cases of fatality. In April 2009, a new strain of human influenza A (H1N1) virus containing genes from human, swine and avian influenza A viruses emerged in Mexico. Over the course of one year, the virus had spread to more than 212 countries and overseas territories or communities, causing more than 15 921 deaths. More recently, in January 2012, a human case of avian influenza A (H5N1) virus infection was reported in China. If history is our guide, we can assume that the threat of these diseases will continue to grow and pose a serious problem to the security of countries worldwide.

7.2 Microbial sequence analysis

7.2.1 Sequence alignment

Similarity between related sequences can give clues to the structure, function, or homology to the common ancestor. Computational methods that can compare sequence features are, therefore, particularly useful. Sequence alignment is the determination of residue–residue correspondences between two or more character strings, usually preserving the relative order. This method allows us to measure similarity and infer
evolutionary relationships between two or more sequences. Pairwise sequence alignment is useful for analyzing the degree of similarity between two biological sequences. Where more than two sequences are involved, multiple sequence alignment can be used to identify regions of similarity that may help explain functional and/or phenotypic variability. The 2009 H1N1 flu was not the first human pandemic caused by influenza A viruses. It is related to the 1889 Russian flu that killed ~1 million people, the 1918 Spanish flu that infected ~25% of the global population and killed at least 50 million people worldwide, the 1957 Asian flu that resulted in ~2 million deaths and the 1968 Hong Kong flu that caused ~1 million deaths. In cases where the ancestry is unclear, sequence alignment methods can be used to infer their phylogenetic relationships. This includes:

- identifying globally optimal alignment solutions for studying highly conserved sequences;
- identifying maximally homologous subsequences among sets of long sequences for detecting distantly related proteins.

In information theory and computer science, four types of metrics are commonly used to measure the edit distance between two strings of characters. They include:

- The Hamming distance, which is the number of positions with mismatched characters between two strings of the same length.
- The Levenshtein distance, which is the minimum number of operations that is needed to transform one string into the other, which may be of different length. An operation can be a deletion, insertion, or substitution of a single character in the strings.
The Damerau-Levenshtein distance, which is the minimum number of operations that is needed to transform one string into the other, which may be of different length. An operation can be a deletion, insertion, or substitution of a single character, or a transposition of two adjacent characters in the strings.

The Jaro-Winkler distance, which is a measure of similarity between two strings using the Jaro distance metric. This method first identifies the common characters between two strings of characters. Two characters are common if there is an exact match and if the difference in positions between the two strings is less than half the length of the shorter string. Once all the common characters are determined, the number of transpositions of common characters are determined and used to compute the Jaro similarity. Strings that are more similar will have a higher Jaro distance.

In biological systems, certain amino acid changes are more likely to occur than others. For example, a hydrophobic residue is more likely to be replaced by another hydrophobic residue than a hydrophilic residue. To account for such transformations, a weight can be assigned to the different edit operations. This can take the form of a matrix that shows the substitution frequencies of observed pairs of amino acid residues. Two popular substitution matrices are:

The Percent Accepted Mutation (PAM) matrices by Dayhoff, which measure sequence similarity in closely related species. Two sequences 1 PAM apart have an average of one accepted point mutation event per 100 amino acids. They need not be 99% identical, as two point accepted mutations can occur at the same position. To analyze sequences that are more divergent, we can use the PAM1 matrix as a base for calculating other matrices.
This is based on the assumption that repeated mutations would follow the same pattern as those in the PAM1 matrix.

- The BLOck SUbstitution Matrix (BLOSUM) matrices by Henikoff and Henikoff, which measure sequence similarity in divergent sequences. The matrices are constructed from the BLOCKS database of aligned conserved regions in divergent protein families. These regions are assumed to be of functional importance.

Once the substitution matrix is selected, the optimal alignment can be found using dynamic programming algorithms.

### 7.2.2 Information entropy

A related concept is the use of theoretical statistics, such as information entropy, to quantify the rate of information transfer in biological sequences. The Shannon entropy is a measure of uncertainty that is associated with a random variable. It is commonly used to assess the variability of microbial proteomes and epitope sequences. For a given alignment, the information content (i.e. entropy) of an amino acid position $H(x)$ is defined by:

$$H(x) = -\sum P(x) \log P(x)$$

where $x$ is one of 20 amino acid residue types. $P(x)$, the probability of occurrence of $x$, is estimated by $f(x)$, the frequency of the appearance of residue type within the alignment column:

$$P(x) = f(x) = N(x)/L$$

where $N(x)$ is the number of appearances of amino acid residue $x$, and $L$ is the length of the column. This method has
been used to analyze the genetic diversity and antigenic relationships of Chikungunya virus (CHIKV) proteomes from its introduction in 1952 to 2009. Antigenic switches refer to changes in gene expression at a specific site which may abrogate binding to HLA molecules or antagonize/interfere with T cell response, leading to cellular immune evasion. The study suggested that CHIKV is undergoing mild positive selection, with significant amounts of “antigenic switches” clustered over the entire genome.

An effective way to identify amino acid residues that are involved in virus adaptation is to find interdependencies between mutations in multiple proteins. A simple way to do this is to calculate mutual information (MI) between variable pairs. MI is an information theoretical statistic that measures the strength of association between a pair of variables. The mutual information between two variables A and B is defined by:

\[ MI(A, B) = H(A) + H(B) - H(A, B) \]

where \( H(A) \) and \( H(B) \) are entropies of A and B, while \( H(A, B) \) is their joint entropy over a set of all possible unique pairs of values \((A, B)\). This method can also be used to identify characteristic sites in pathogen proteins with conserved mutations. For example, Miotto and colleagues applied this method in 2010 to identify adaptive signatures of influenza A proteomes from 92 343 sequences.

### 7.3 Detecting natural selection in molecular evolution

The evolutionary inertia of a pathogen can be qualitatively examined by studying the nucleotide usage patterns at single amino acid sites. The neutral theory of molecular evolution
by Kimura in 1968 states that most evolutionary changes at the molecular level are caused by random genetic drift of selectively neutral nucleotide substitutions. Due to the degeneracy of the genetic code, some point mutations are silent with no amino acid replacements. Silent or synonymous substitutions are primarily transparent to natural selection, whereas replacement or non-synonymous substitutions may be a result of strong selective pressure. A simple method to calculate the extent of adaptive evolution at highly variable genetic loci is to compare the fixation rates between non-synonymous (\(d_N\)) and synonymous (\(d_S\)) substitutions.

The \(d_N/d_S\) ratio (\(\omega\)), otherwise known as the “acceptance rate,” provides a sensitive measure of selection pressure at the amino acid level. \(\omega=1\) indicates neutral expectation, \(\omega<1\) suggests negative (purifying) selection, while \(\omega>1\) suggests positive (diversifying) selection. A group of genes that often show the \(\omega>1\) relationship are antigenic genes in human immunodeficiency virus-1, plasmodia, and other parasites. The hemagglutinin gene from influenza A virus is probably one of the fastest evolving genes in terms of the rate of nucleotide substitution, which was estimated at 5.7×10\(^{-3}\) per site per year. This high genetic variation confers a fitness advantage to the pathogen in its attempt to evade host defenses.

The simple counting method of Nei and Gojobori is commonly used for estimating \(d_N\) and \(d_S\). However, the reliability of this technique is low when the rate of transitional nucleotide change is higher than that of transversional change. The model-based maximum likelihood (ML) methods such as those proposed by Muse and Gaut and Goldman and Yang represent a viable and widely used alternative for this purpose. The original ML model of Goldman and Yang assumes a single \(\omega\) for all lineages and sites, and has been extended to account for variation by
allowing $\omega$ to vary either across lineages, among substitution sites, or both among sites and among lineages. Lineage-specific models assume that $\omega$ do not vary among sites, and can detect positive selection for a lineage only if the averaged $d_N$ over all sites is greater than the average $d_S$. Site-specific models, on the other hand, allow $\omega$ to vary among sites but not among lineages. As such, these models can detect positive selection at individual sites only if the averaged $d_N$ over all lineages is greater than the average $d_S$. By allowing $\omega$ to vary both among sites and among lineages, the method can be applied to detect positive selection that occurred at a few time points and affects a few sites.

### 7.4 Bibliography

1. Kemena, C. and Notredame, C. (2009) Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics* 25: 2455–65.
2. Bhattacharya, T., Daniels, M., Heckerman, D., Foley, B., Frahm, N., et al. (2007) Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science* 315: 1583–6.
3. Shannon, C.E. (1950) Prediction and entropy of printed English. *Bell System Technical Journal* 30, 50–64.
4. Tong, J.C., Simarmata, D., Lin, R.T.P., Renia, L. and Ng, L.F.P. (2010) HLA class I restriction as a possible driving force for Chikungunya evolution. *PLoS ONE* 5: e9291.
5. Miotto, O., Heiny, A.T., Albrecht, R., Garcia-Sastre, A., Tan, T.W., et al. (2010) Complete-proteome mapping of human influenza A adaptive mutations: implications for human transmissibility of zoonotic strains. *PLoS ONE* 5: e9025.
6. Sheng, C., Hsu, W., Lee, M.L., Tong, J.C. and Ng, S.K. (2010) Mining mutation chains in biological sequences. *Proc. Int. Conf. on Data Engineering ICDE*: 473–84.

7. Grenfell, B.T., Pybus, O.G., Gog, J.R., Wood, J.L.N., Daly, J.M., et al. (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**: 327–32.

8. Bennett, S.N., Holmes, E.C., Chirivella, M., Rodriguez, D.M., Beltran, M., et al. (2003) Selection-driven evolution of emergent dengue virus. *Mol. Biol. Evol.* **20**: 1650–8.

9. Suzuki, Y. and Gojobori, T. (1999) A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* **16**: 1315–28.

10. Suzuki, Y., Gojobori, T. and Nei, M. (2001) ADAPTSITE: detecting natural selection at single amino acid sites. *Bioinformatics* **17**: 660–1.

11. Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature* **217**: 624–6.

12. Nei, M. (2005) Selectionism and neutralism in molecular evolution. *Mol. Biol. Evol.* **22**: 2318–42.

13. Miyata, T. and Yasunaga, T. (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its applications. *J. Mol. Evol.* **16**: 23–36.

14. Air, G.M. (1981) Sequence relationships among the hemagglutinin genes of 12 subtypes of influenza A virus. *Proc. Natl. Acad. Sci. USA* **78**: 7639–43.

15. Nei, M. and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**: 418–26.
16. Muse, S.V. and Gaut, B.S. (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* 11: 715–24.
17. Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11: 725–36.
18. Kosakovsky Pond, S.L., Poon, A.F.Y., Brown, A.J.L. and Frost, S.D.W. (2008) A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus. *Mol. Biol. Evol.* 25: 1809–4.
19. Yang, Z. and Nielsen, R. (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19: 908–17.
