Anisimova *BMC Evolutionary Biology* (2015) 15:76
DOI 10.1186/s12862-015-0352-y

BMC Evolutionary Biology

REVIEW Open Access

Darwin and Fisher meet at biotech: on the potential of computational molecular evolution in industry

Maria Anisimova[1,2,3]

Abstract

Background: Today computational molecular evolution is a vibrant research field that benefits from the availability of large and complex new generation sequencing data – ranging from full genomes and proteomes to microbiomes, metabolomes and epigenomes. The grounds for this progress were established long before the discovery of the DNA structure. Specifically, Darwin's theory of evolution by means of natural selection not only remains relevant today, but also provides a solid basis for computational research with a variety of applications. But a long-term progress in biology was ensured by the mathematical sciences, as exemplified by Sir R. Fisher in early 20th century. Now this is true more than ever: The data size and its complexity require biologists to work in close collaboration with experts in computational sciences, modeling and statistics.

Results: Natural selection drives function conservation and adaptation to emerging pathogens or new environments; selection plays key role in immune and resistance systems. Here I focus on computational methods for evaluating selection in molecular sequences, and argue that they have a high potential for applications. Pharma and biotech industries can successfully use this potential, and should take the initiative to enhance their research and development with state of the art bioinformatics approaches.

Conclusions: This review provides a quick guide to the current computational approaches that apply the evolutionary principles of natural selection to real life problems – from drug target validation, vaccine design and protein engineering to applications in agriculture, ecology and conservation.

Keywords: Molecular evolution, Applied bioinformatics, Modeling, Selection, Adaptation, Conservation, Drug target, Resistance, Immune response

Introduction

For over a century computational scientists have been working side by side with empirical scientists, supporting key developments in molecular and evolutionary biology. Despite this, today close interdisciplinary collaboration can be still somewhat elusive, with different communities of scientists speaking "different languages". Yet, it is well worth adapting the research process and communication in order to include a wider range of specialists, particularly in industries.

A historical perspective shows that progress in life sciences relies on solid backing from statisticians, mathematicians, computational scientists, and theoreticians in general. Remarkable in this context is the contribution by R. A. Fisher – one of the first bioinformaticians, who developed the statistical theory for experimental design and hypothesis testing, together with many now widely used techniques (eg, the analysis of variance, the method of maximum likelihood, etc.), originally to address the needs of agricultural research at the Rothamsted Experimental Station, Together with S. Wright and J. B. S. Haldane, Fisher has established the field of population genetics, and contributed to the neo-Darwinian evolutionary synthesis, which reconciled Mendelian genetics with Darwin's evolutionary theory at the level of hereditary molecular information. In the 60s the founders of molecular

Correspondence: maria.anisimova@zhaw.ch
[1]Institute of Applied Simulations, School of Life Sciences and Facility Management, Zürich University of Applied Sciences, Einsiedlerstrasse 31a, Wädenswil 8820, Switzerland
[2]Department of Computer Science, ETH, Zurich, Switzerland
Full list of author information is available at the end of the article

evolution E. Zuckerkandl and L. Pauling used quantitative comparisons to show that molecular changes in a protein accumulate at a uniform rate [1]. This concept, known as "molecular clock", enabled the theoretical work by J. Crow and M. Kimura, who modeled genetic drift and selection as realizations of similar processes. The molecular clock served as basis for Kimura's neutral theory of molecular evolution, whereby selection had no significant influence on shaping genomes with most genetic changes being selectively neutral [2]. The neutral theory greatly contributed to the development of the field as it provided a simple null hypothesis with testable predictions. Since then, numerous statistical tests have been developed and remain highly relevant to detecting selection in genomic data, as emphasized later in this review.

More recently, the availability of high-throughput molecular data served to advance statistical and computational methods for genomics, allowing for a variety of applications – from medical genetics and pharmacology to biotechnology, agriculture and ecology. The size and the complexity of molecular data underline the crucial role of theoreticians and computational scientists for the success of biological data exploration . Molecular data size and complexity have surpassed the so-called "Excel barrier", so that companies analyzing genomics data can no longer rely on old practices. Indeed, pharma and biotechnology companies see an increasing demand for computational scientists with strong skills in mathematical modeling, machine learning, data mining, complex optimization and data representation (e.g., [3]).

Review

The importance of selection studies at the genomic level

The field of computational genomics has been growing steadily, attracting more research funding for both academic and applied research in biotech and pharma companies. Here I focus on the potential of computational methods to study how genomic changes occur over time and their impact on phenotype or genetic fitness [4-6]. While Darwin has described how selection may act on a phenotype, he had no knowledge of hereditary mechanisms, and would have been pleased to see how far we have come today in our understanding of selective mechanisms in molecular sequences. Current computational methods can detect genomic regions under selection and help to describe the biological mechanisms generating the observed molecular patterns. Considering this, computational methods provide effective means of narrowing down the space of plausible candidates or hypotheses for further testing. A diversity of biological mechanisms may cause genetic mutations with various fitness effects, leading to a variety of ways natural selection can manifest itself. The central role of selection on molecular sequences has been demonstrated in the adaptation to new environments, the host-pathogen "arms" race, the emergence of competition, the evolution of complexity, and in the morphological and behavioral evolution, for example see Figure 1 of [7]. Natural selection may act on the protein, on the DNA sequence, and even on whole genomic features. Negative or purifying selection conserves the sequence (or other molecular features), while positive selection acts in a diversifying or a directional manner favoring specific changes. Positive selection typically affects molecular regions involved in genetic conflict, and often acts in an episodic manner (i.e., for a limited time). Selection scans became an indispensible component of genomic studies (e.g., [8,9]), since they help to understand the biological constraints and to identify mutational hotspots due to adaptive processes.

Studies of selective constraints in genomes of populations and species can have a variety of applications (see Table 1 for examples). Identification of deleterious mutations (e.g., mutations causing disease) may aid the development of gene therapies and personalized treatments. Detecting hotspots of diversifying pressure in antigenic sites, epitopes and pathogenic receptors can be used in drug and vaccine design. Phylogenetic methods are increasingly used in immunology and cancer genomics. The analysis of selective pressures and disease transmission rates using host and pathogen samples provides important clues for epidemiology, helping to understand the disease dynamics and to develop predictive strategies for disease control. This applies equally to animal and plant hosts as well as their pathogens, thus having applications also in the domain of agricultural research such as developing molecular-based strategies for increasing crop resistance to pathogens. Similarly, evolutionary studies may provide insights to the genetic basis for stress tolerance and yields of animal and plant products. Other applications of molecular evolution and selection analyses may include biodiversity, conservation, sustainable development, bioremediation, bioengineering and nutrition. Below I briefly draw attention to some successful approaches for studying the evolutionary dynamics in molecular sequences, illustrated by examples.

Computational approaches to study evolution and selection in molecular sequences

Evaluating selective pressures on molecular sequences relies on the comparative evolutionary approach, and therefore requires at least two homologous sequences [10]. Simple studies of sequence conservation already go towards this objective – they allow to pinpoint functionally important parts of a sequence, based on our understanding of how natural selection acts on molecular sequences. In practice, studies of sequence homology and conservation have been fundamental to the discovery in genomics. In pharma industry, alignment and

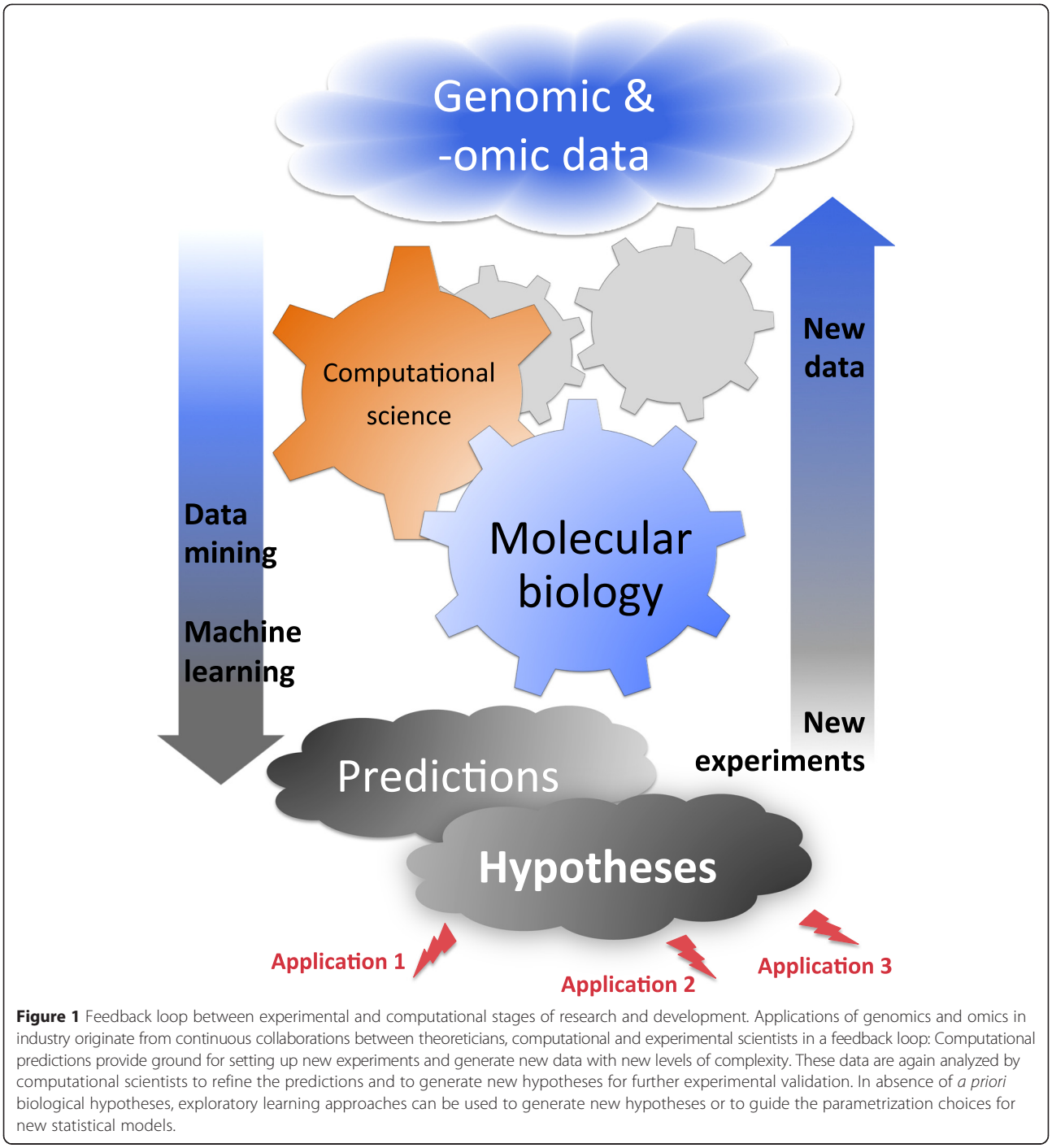

Figure 1 Feedback loop between experimental and computational stages of research and development. Applications of genomics and omics in industry originate from continuous collaborations between theoreticians, computational and experimental scientists in a feedback loop: Computational predictions provide ground for setting up new experiments and generate new data with new levels of complexity. These data are again analyzed by computational scientists to refine the predictions and to generate new hypotheses for further experimental validation. In absence of *a priori* biological hypotheses, exploratory learning approaches can be used to generate new hypotheses or to guide the parametrization choices for new statistical models.

similarity searches are routinely used together with template-based structure prediction in structure-based drug design (integrated in purpose-built software).

For protein-coding genes, codon models of substitution provide means to expand inferences from simple sequence conservation to more sophisticated modeling of codon substitution through time driven by selection and mutation [11-17]. In such models selection is modeled explicitly, allowing for variation of selection pressure across sequence sites and over time. The power of the approach depends on the number and the range of sequences analyzed [18]. For large samples from well-designed experiments, it is possible to accurately predict the positions and the time episodes where selection has operated [19-22]. Other tests for selection are not specific to coding sequences (for review see [7]). Substitution models in general can be used to detect shifts in evolutionary rates or sequence composition [23-25],

Table 1 Selected examples of applications of molecular evolution and selection studies

Application type	Description	Citation	Computational approach
Control of HIV infection	Protein function study of HIV restriction properties in TRIM5α	[40]	Codon model tests for selection
Model species selection for pharmaceutical discovery	Assessment of pharmacological target homology	[42]	Phylogenetic analyses of gene families
HIV vaccine development	Assessment of phylogenetic diversity in viral proteins and antibodies; identification of conserved epitopes	[50,53]	Phylogenetic analyses and codon model tests for selection
Flu epidemics prediction; vaccine strain selection	Modeling of antigenic dynamics of flu over time	[54]	Phylogenetic diffusion model of antigenic evolution
Prediction of HIV progression	Monitoring the synonymous substitution rates in viral protein samples from HIV-positive patients over time	[67]	"Relaxed-clock" modeling of codon evolution
Evaluating epidemics dynamics and the effect of public health interventions	Estimating the rates of transmission, recovery, sampling, and the effective reproductive number	[81-83]	Birth-death phylogenetic models
Flu epidemics prediction; vaccine strain selection	Modeling adaptive epitope changes and deleterious mutations outside the epitopes in flu from one year to the next	[93]	Molecular evolution modeling over viral genealogies
Crop resistance	Identifying the resistant variants of the *Pi-ta* gene in rice that is used to control rice blast disease	[96]	Analyses of genetic diversity and evolution
Mapping disease associations; complex disease biology; development personalized medicine	Genome studies identifying sites of genomic diversification, associations with diseases, estimating fitness of mutations	[73,74]	Evolutionary analyses of genomic constraints, genome-wide association studies
***Disease biology; identification of vaccine targets**	Population genomics of the sexually transmitted bacteria *Chlamydia trachomatis*	[97]	Genome-wide evolutionary analyses of conservation by codon models and population genetics approaches
***Disease biology**	Adaptation in the cavity causing bacteria *Streptococcus mutans*	[98]	Genome-wide evolutionary analyses of conservation and demography
***Conservation and biodiversity; climate change**	Evaluating hybridization of blue whale subspecies in southern hemisphere	[99]	Population genetics analyses
***Impact of climate change**	Evaluating the interplay between global climate change, genetic diversity and species interactions and community structure	[100]	Evaluation of intraspecific genetic diversity by population genetics approaches

*Highlighted in the 2013 editorial "Highlights in applied evolutionary biology" in the peer-reviewed journal "Evolutionary Applications".

which are potentially due to adaptive processes. Neutrality tests based on summary statistics allow inferences of selection if no other demographic factors can be invoked to explain the observed data [26,27]. Besides these methods, selection can be detected using Poisson random-field models [28-30], and tests based on linkage disequilibrium, haplotype structure and population differentiation [31-36].

The basic idea behind all tests for selection is to compare the molecular patterns observed in genomic sequences to what could be expected by chance. Significant deviations point to interesting candidate regions, sites or time episodes, and provide excellent hypotheses for further experimental and statistical testing. Different methods use different statistics to make their inferences about selection. Ideally, the null expectation and alternative scenarios can be described by a statistical model. This enables a proper statistical treatment during parameter estimation, evaluation of uncertainty, hypothesis testing and model selection. Model-based approaches, while desirable, should make sure to use models that account for key biological factors and that are sufficiently robust against violations of key assumptions. It is important to be aware, that biological mechanisms that are not included in the model may have a significant impact on the objective of inference. If it is not possible to include a certain biological factor (e.g., population size) in a model, its influence on the parameter of interest (e.g., selection pressure on the protein) can be investigated using separate carefully designed tests.

In this respect, Markov models of character substitution have been particularly successful at inferring selection at individual sites and lineages. Among widely used methods are likelihood ratio tests of codon substitution models, which detect selection on the protein sequence

using the comparison of nonsynonymous (amino-acid altering) and synonymous (amino-acid preserving) substitution rates (for review see [37]). If a test is significant, Bayesian prediction is used to identify the selected positions or lineages affected by selection. The pharmaceutical giant GlaxoSmithKline (GSK) acknowledged the applied value of these methods by an award to the principal investigator Prof Ziheng Yang (UCL, UK). The relevance of selection analyses with codon models for downstream applications can be demonstrated on a selection of case studies. A classic example is the human major histocompatibility complex molecules of class I (glycoproteins mediating cellular immunity against intracellular pathogens), where all residues under diversifying selection pressure were found clustered in the antigen recognition site [38,39]. In another example, selection analyses identified a sequence region of 13 amino acids with many positive-selected sites in TRIM5α, involved in cellular antiviral defense [40]. Functional studies of chimeric TRIM5α genes showed that the detected region was responsible for the difference in function between the rhesus monkey linage where TRIM5α restricts HIV-1 and the human TRIM5α that has only weak restriction.

More generally, the numerous genome-wide scans in mammals agree that genes affected by positive diversifying selection are largely responsible for sensory perception, immunity and defense functions [41]. Consequently, pharma and biotech companies should make a greater use of computational approaches to detect genes and biochemical pathways subject to differential adaptive evolution in human and other lineages used as experimental model organisms, for example as it has been done by R & D of GSK [42,43]. Such studies can be extremely valuable, for example when selecting drug targets. Particularly, evolutionary analyses can pinpoint evolutionary differences between model organisms used for drug target selection. Such differences can be responsible for unpredicted disparities in response to medical treatment, as it has been highlighted by the tragic effects of TGN1412 treatment during human drug trials in 2006 [44]. Selection analyses are also important for research in agriculture or conservation, since in plant genomes positive selection affects most notably disease resistance genes [45,46], defense enzymes such as chitinases [47] and genes responsible for stress tolerance [48]. Consequently evolutionary studies help to detect proteins, binding sites and their interactions relevant for host-pathogen coevolution. For example, diversifying selection drives the evolution of several exposed residues in leucine-rich repeats (LRRs) of the bacterial type III effectors (that attack plant defense system) from the phytopathogenic *R. Solancearum* infecting >200 of plant varieties including agriculturally important crops [49]. Similarly, studies of phylogenetic diversity and selection in viral strains and antibody sequences are contributing to the new HIV vaccine development strategy, whereby antibodies are designed to bind to conserved epitopes of selected viral targets [50-53]. Moreover, molecular evolution modeling approaches can greatly enhance the modeling of antigenic dynamics of pathogens over time (e.g., [54]).

In protein coding sequences selection may also act on the DNA, whereby synonymous codon changes may affect protein's stability, expression, structure and function [55,56]. Translational selection manifests itself as the overall codon bias in a gene to match the abundances of cognate tRNA. Remarkably, this property can be successfully used in biotechnology, for example to dramatically increase transgene expression by synthesizing sequences with optimal synonymous codons [57]. Optimal codon usage may be approximated by codon usage bias – using bioinformatics methods [58]. Besides this, more subtle selective mechanisms may act on certain codon positions; affecting splicing, mRNA stability, gene regulation protein abundance, folding and function (e.g., [59-63]). In human genes this may lead to disease (such as cancers and diabetes) or may be responsible for differences in individual responses to drug treatment (e.g., [64]). Haplotypes with synonymous changes may have increased fitness and will be consequently increase in frequency in a population. Therefore, the knowledge of these specific synonymous polymorphisms may be important to explain differential treatment effects in population and contribute to the development of personalized medicines [65]. Molecular evolution methods are powerful enough to detect such interesting candidate cases: Recent study of synonymous rates detected many disease related genes, particularly associated with various cancers, as well as many metabolizing enzymes and transporters, which affect the disposition, safety and efficacy of small molecule drugs in pharmacogenetics [66]. This shows that computational molecular evolution studies have real power to predict genes and codon positions where a replacement of synonymous codons changes protein fitness. Such predictions promise to be valuable for applications in protein engineering. Indeed, some biotech companies such as DAPCEL are already using the knowledge of interesting synonymous positions for enhanced protein production. Compared to laborious and time-consuming trial-and error experiments, computational prediction offers a fast way of obtaining candidate genes and positions for experimental validation. Furthermore, monitoring of the synonymous rates may be also informative for diagnostics purposes, as has been shown in evolutionary studies of serial viral samples from HIV-positive patients [67].

Species evolution is however a result of complex population dynamics, making population scale studies of genetic diversity a powerful complement to codon-based selection analyses. Successful population level techniques

include tests of neutrality [26], Poisson random-field models (e.g., [68,69]) combined with demographic modeling and genome-wide association studies [70]. These methods apply to full genome sequences helping to identify also non-coding genomic regions of functional relevance and those associated with certain population traits. For medical genetics, uncovering the relevance of genomic variation in populations helps to pinpoint the disease variants and use this information in the development of personalized medicines and treatments. Determining fitness of specific mutations is now possible using macroevolutionary inferences and population genetics approaches [71-73], which can be successfully combined with genome-wide association studies [74]. These inferences could be combined with applications in a clinical context [75].

However, many traits are shaped by multiple loci so that the effects of any single mutation can be observed only through their epistatic effects [76,77]. Consequently, computational approaches recently extended single loci inferences to detecting epistatic effects of mutations through the identification of polygenic selection, i.e., whereby selection affects whole gene clusters whose protein products interconnected in the biological pathways that they share. Such analyses found that polygenic selection often affects pathways involved in immune response and adaptation to pathogens [78], which is also consistent with results from single loci studies.

Another approach for detecting selective signatures is based on detecting shifts in evolutionary substitution rates over time, for example based on covarion or Markov modulated models [16,79]. Such methods may be used to detect functional shifts in proteins of interest, providing evolutionary information that aids structural and functional protein prediction. Therefore such analyses can be helpful for many pharma and biotech applications that use structural modeling to design proteins and peptides for therapeutic or other biotechnology applications (e.g., [80]). Alternatively, changing diversification rates can provide evidence for changing environments, emerging pathogens and shed light on epidemiological dynamics. Diversification bursts or exponential growth, for example, may represent the emergence of particularly virulent strains resulting in epidemics. Such selective signatures can be characterized by phylogenies or genealogies relating the molecular sequences in a viral sample based on the birth-death models of stochastic branching processes [81]. This approach allows to evaluate the effects of public health interventions by estimating the rates of transmission, recovery, and sampling, and consequently, the effective reproductive number. For epidemiology-related problems, these techniques become particularly powerful when combined with classical epidemiologic models SIR or SIS [82,83]. Evolutionary methods can be useful also for the analyses of somatic hypermutation in antibody sequences during antibody maturation, or for monitoring somatic mutations in cancerous tissues [84-87]. Indeed, applications of phylogenetic methods to cancer and immunology research are now attracting more attention and funding (e.g., [88,89]).

Selection may also operate on whole genomic features, such as indels, gene order, gene copy numbers, transposable elements, miRNAs, post-translational modifications, etc. To detect selective signatures of conservation or adaptation, the observed genomic patterns are compared with a neutral expectation, i.e., patterns that can arise by chance alone. For example, phylogenetic patterns produced by tandem repeats in eukaryotic proteins can be used to identify interesting candidate genes that might be under diversifying pressures [90]. In plants a similar analysis strongly pointed to lineages where diversification (in terms of unit number and their order conservation) occurs in LRRs that are found in abundance in plant resistance genes [91]. Such analyses allow for example to pinpoint the relevant genes and lineages where selection on tandem repeat units is due to adaptation to emerging pathogens or to changing environmental conditions. This opens the door to applications such as synthetically introducing identified gene variants into plant genomes to produce crops with improved resistance or better stress tolerance properties. Indeed, crop protection agencies and companies (e.g., Syngenta, Rothamsted Research) have started using evolutionary analyses to elucidate the origins of resistance to pathogens [92].

Even when selection study is not the goal of the analyses, modeling its influence on genomic data is of utmost importance. Failing to do so may lead to biased and inaccurate inferences that could misguide follow-up experimental studies. However, modeling selection enhances the predictive power of methods that are used to study adaptive or antagonistic processes. A nice example is the recent predictive fitness model for influenza, which couples the fitness values and frequencies of strains with molecular evolution modeling on an influenza stain genealogy for haemaglutinin gene [93]. This approach uses observed viral samples taken from year to year to predict evolutionary flu dynamics in the coming year, which is practically relevant for selecting vaccine strains for the new flu season.

Conclusions

The last decades have seen the development of accurate and powerful computational methods for evaluating evolution and selection in molecular data. Both industry and basic research should discover and exploit the full potential of these methods, as they provide efficient means to generate viable biological hypotheses, including interesting candidates cases for further experimental

testing. Examples above included such successful applications. Genomics and omics data provides immense opportunities for applications in industry. While pharma and biotech giants are generally aware of this, employing their own bioinformaticians, smaller companies often do not know of current possibilities provided by the state-of the art computational methods and the volumes of newly generated data. But even bioinformatics teams in pharmaceutical giants usually have no sufficient capacity to develop suitable techniques for the analysis of their data, as they focus on more imminent results for their company. The development of robust statistical methodology demands substantial time doing basic research. Further, even for method developers it is hard to keep pace with all the relevant advances in the field. For this reason, industry should actively engage with researchers in academia – starting with networking and discussions of company's needs and the potentially useful academic results, gradually bringing this into productive industry-academia collaborations.

Finally, productive collaborations require efficient communication between theoreticians, computational and experimental scientists, with a continued feedback loop built into the research process (Figure 1). While research in biology is traditionally hypothesis-driven, current volumes of new complex data also require exploratory learning approaches. Therefore today, machine learning, pattern recognition and data mining approaches became essential in the exploration of big data. Such techniques can help with the choice of suitable model parameterizations [94]. Computational predictions can be used to formulate new biological hypotheses. To validate these hypotheses, new experiments should be set up in order to generate new data, possibly with new levels of complexity. These data can be analyzed again by computational scientists, in order to refine the initial predictions and to refine or generate new hypotheses for further experimental validation, re-starting the loop (Figure 1). While it is easy to generate genomic data today, greater thought must be invested into the experimental design, in order to make the statistical inferences more accurate and informative. This requires solid expertise in statistics.

To conclude, that pharma and biotech companies should actively seize upon the potential of computational molecular evolution approaches in their translational research. As shown above, this can include drug target identification and validation, animal model selection, preclinical safety assessment, vaccine design, epidemics control and drug repositioning. Such techniques are promising to become mainstream, strengthening the current position of translational research in industry. The translational value of computational molecular evolution is not limited to health and pharma industry, but also include a variety of other exciting applications – protein engineering, agriculture, environmental risk assessment, ecology, biodiversity and conservation. Again, this cannot be done without strong interdisciplinary partnerships. Bioinformatics has now become a vibrant and highly interdisciplinary area of research and the outlook for its future and its applications is very optimistic – "Bioinformatics alive and kicking" [95].

Competing interests
The author declares that she has no competing interests.

Authors' contributions
The author has conceived the idea and written this article.

Acknowledgements
This article was motivated by the mini-symposium "Life in numbers: the use of modeling and simulation in life sciences" which took place on the 5th of June 2014 at the Zürich University of Applied Sciences (http://2014.lifeinnumbers.ch) in partnership with the Swiss Institute of Bioinformatics and sponsored by Actelion, Philips and Geneious (Biomatters). The event intended to highlight the potential of computational biology in the applied domain and to bring together scientists from academia and industries. The first version of this article forms part of the Festschrift collection for the GNOME 2014 symposium to celebrate the retirement of Prof. Gaston Gonnet (ETH Zürich), and as such has been published open access as a pre-print with PeerJ.

Author details
[1]Institute of Applied Simulations, School of Life Sciences and Facility Management, Zürich University of Applied Sciences, Einsiedlerstrasse 31a, Wädenswil 8820, Switzerland. [2]Department of Computer Science, ETH, Zurich, Switzerland. [3]Swiss Institute of Bioinformatics, Lausanne, Switzerland.

Received: 2 July 2014 Accepted: 15 April 2015
Published online: 01 May 2015

References
1. Zuckerkandl E, Pauling L. Molecules as documents of evolutionary history. J Theor Biol. 1965;8(2):357–66.
2. Kimura M. Evolutionary rate at the molecular level. Nature. 1968;217(5129):624–6.
3. Price M. Computational biologists: the next pharma scientists? Science Careers. 2012 (April 13).
4. Yang Z. Computational molecular evolution. UK: Oxford University Press; 2006.
5. Cannarozzi GM, Schneider A. Codon evolution: mechanisms and models. UK: Oxford University Press; 2012.
6. Anisimova M. Evolutionary genomics: statistical and computational methods. New York: Humana press, Springer; 2012.
7. Anisimova M, Liberles D. Detecting and understanding natural selection. In: Cannarozzi G, Schneider A, editors. Codon evolution: mechanisms and models. Oxford: Oxford University Press; 2012.
8. Fu W, Akey JM. Selection and adaptation in the human genome. Annu Rev Genomics Hum Genet. 2013;14:467–89.
9. Stapley J, Reger J, Feulner PG, Smadja C, Galindo J, Ekblom R, et al. Adaptation genomics: the next generation. Trends Ecol Evol. 2010;25(12):705–12.
10. Nielsen R, Hubisz MJ. Evolutionary genomics: detecting selection needs comparative data. Nature. 2005;433(7023):E6. discussion E7–8.
11. Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 1994;11(5):725–36.
12. Yang Z, Nielsen R, Goldman N, Pedersen A-MK. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 2000;155(1):431–49.
13. Yang Z, Nielsen R. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol. 2008;25(3):568–79.
14. Muse SV, Gaut BS. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol. 1994;11(5):715–24.

15. Kosakovsky Pond SL, Muse SV. Site-to-site variation of synonymous substitution rates. Mol Biol Evol. 2005;22(12):2375–85.
16. Guindon S, Rodrigo AG, Dyer KA, Huelsenbeck JP. Modeling the site-specific variation of selection patterns along lineages. Proc Natl Acad Sci U S A. 2004;101(35):12957–62.
17. Rodrigue N, Philippe H, Lartillot N. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. Proc Natl Acad Sci U S A. 2010;107(10):4629–34.
18. Anisimova M, Bielawski JP, Yang Z. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol. 2001;18(8):1585–92.
19. Anisimova M, Bielawski JP, Yang Z. Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol. 2002;19(6):950–8.
20. Anisimova M, Yang Z. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol. 2007;24(5):1219–28.
21. Lu A, Guindon S. Performance of standard and stochastic branch-site models for detecting positive selection among coding sequences. Mol Biol Evol. 2014;31(2):484–95.
22. Yang Z, dos Reis M. Statistical properties of the branch-site test of positive selection. Mol Biol Evol. 2011;28:1217–28.
23. Galtier N, Jean-Marie A. Markov-modulated Markov chains and the covarion process of molecular evolution. J Comput Biol. 2004;11(4):727–33.
24. Wang HC, Spencer M, Susko E, Roger AJ. Testing for covarion-like evolution in protein sequences. Mol Biol Evol. 2007;24(1):294–305.
25. Blanquart S, Lartillot N. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Mol Biol Evol. 2006;23(11):2058–71.
26. Nielsen R. Statistical tests of selective neutrality in the age of genomics. Heredity. 2001;86(Pt 6):641–7.
27. Zhen Y, Andolfatto P. Methods to detect selection on noncoding DNA. Methods Mol Biol. 2012;856:141–59.
28. Bustamante CD, Wakeley J, Sawyer S, Hartl DL. Directional selection and the site-frequency spectrum. Genetics. 2001;159(4):1779–88.
29. Zhu L, Bustamante CD. A composite-likelihood approach for detecting directional selection from DNA sequence data. Genetics. 2005;170(3):1411–21.
30. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. Localizing recent adaptive evolution in the human genome. PLoS Genet. 2007;3(6):e90.
31. Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andres AM, Albrechtsen A, et al. Darwinian and demographic forces affecting human protein coding genes. Genome Res. 2009;19(5):838–49.
32. Chen H, Patterson N, Reich D. Population differentiation as a test for selective sweeps. Genome Res. 2010;20(3):393–402.
33. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006;4(3):e72.
34. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002;419(6909):832–7.
35. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007;449(7164):913–8.
36. Wang ET, Kodama G, Baldi P, Moyzis RK. Global landscape of recent inferred Darwinian selection for Homo sapiens. Proc Natl Acad Sci U S A. 2006;103(1):135–40.
37. Kosiol C, Anisimova M. Selection on the protein-coding genome. Methods Mol Biol. 2012;856:113–40.
38. Hughes AL, Nei M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature. 1988;335(6186):167–70.
39. Yang Z, Swanson WJ. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol Biol Evol. 2002;19(1):49–57.
40. Sawyer SL, Wu LI, Emerman M, Malik HS. Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain. Proc Natl Acad Sci U S A. 2005;102(8):2832–7.
41. Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, et al. Patterns of positive selection in six Mammalian genomes. PLoS Genet. 2008;4(8):e1000144.
42. Vamathevan JJ, Hall MD, Hasan S, Woollard PM, Xu M, Yang Y, et al. Minipig and beagle animal model genomes aid species selection in pharmaceutical discovery and development. Toxicol Appl Pharmacol. 2013;270(2):149–57.
43. Vamathevan JJ, Hasan S, Emes RD, Amrine-Madsen H, Rajagopalan D, Topp SD, et al. The role of positive selection in determining the molecular cause of species differences in disease. BMC Evol Biol. 2008;8:273.
44. Stebbings R, Poole S, Thorpe R. Safety of biologics, lessons learnt from TGN1412. Curr Opin Biotechnol. 2009;20(6):673–7.
45. Meyers BC, Shen KA, Rohani P, Gaut BS, Michelmore RW. Receptor-like genes in the major resistance locus of lettuce are subject to divergent selection. Plant Cell. 1998;10(11):1833–46.
46. Mondragon-Palomino M, Meyers BC, Michelmore RW, Gaut BS. Patterns of positive selection in the complete NBS-LRR gene family of Arabidopsis thaliana. Genome Res. 2002;12(9):1305–15.
47. Bishop JG, Dean AM, Mitchell-Olds T. Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution. Proc Natl Acad Sci U S A. 2000;97(10):5322–7.
48. Roth C, Liberles DA. A systematic search for positive selection in higher plants (Embryophytes). BMC Plant Biol. 2006;6:12.
49. Kajava AV, Anisimova M, Peeters N. Origin and evolution of GALA-LRR, a new member of the CC-LRR Subfamily: from plants to bacteria? PLoS One. 2008;3(2):e1694.
50. Klein F, Mouquet H, Dosenovic P, Scheid JF, Scharf L, Nussenzweig MC. Antibodies in HIV-1 vaccine development and therapy. Science. 2013;341(6151):1199–204.
51. Mouquet H, Klein F, Scheid JF, Warncke M, Pietzsch J, Oliveira TY, et al. Memory B cell antibodies to HIV-1 gp140 cloned from individuals infected with clade A and B viruses. PLoS One. 2011;6(9):e24078.
52. Scheid JF, Mouquet H, Ueberheide B, Diskin R, Klein F, Oliveira TY, et al. Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. Science. 2011;333(6049):1633–7.
53. de Oliveira T, Salemi M, Gordon M, Vandamme AM, van Rensburg EJ, Engelbrecht S, et al. Mapping sites of positive selection and amino acid diversification in the HIV genome: an alternative approach to vaccine design? Genetics. 2004;167(3):1047–58.
54. Bedford T, Suchard MA, Lemey P, Dudas G, Gregory V, Hay AJ, et al. Integrating influenza antigenic dynamics with molecular evolution. eLife. 2014;3:e01914.
55. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 2011;12(1):32–42.
56. Komar AA. Genetics. SNPs, silent but not invisible. Science. 2007;315(5811):466–7.
57. Gustafsson C, Govindarajan S, Minshull J. Codon bias and heterologous protein expression. Trends Biotechnol. 2004;22(7):346–53.
58. Roth A, Anisimova M, Cannarozzi G. Measuring codon-usage bias. In: Cannarozzi G, Schneider A, editors. Codon evolution: mechanisms and models. Oxford: Oxford University Press; 2012.
59. Chamary JV, Parmley JL, Hurst LD. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet. 2006;7(2):98–108.
60. Carlini DB, Genut JE. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. J Mol Evol. 2006;62(1):89–98.
61. Tsai CJ, Sauna ZE, Kimchi-Sarfaty C, Ambudkar SV, Gottesman MM, Nussinov R. Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. J Mol Biol. 2008;383(2):281–91.
62. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, et al. A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science. 2007;315(5811):525–8.
63. Komar AA. A pause for thought along the co-translational folding pathway. Trends Biochem Sci. 2009;34(1):16–24.
64. Sauna ZE, Kimchi-Sarfaty C, Ambudkar SV, Gottesman MM. Silent polymorphisms speak: how they affect pharmacogenomics and the treatment of cancer. Cancer Res. 2007;67(20):9609–12.
65. Fernald GH, Capriotti E, Daneshjou R, Karczewski KJ, Altman RB. Bioinformatics challenges for personalized medicine. Bioinformatics. 2011;27(13):1741–8.
66. Dimitrieva S, Anisimova M. Unraveling patterns of site-to-site synonymous rates variation and associated gene properties of protein domains and families. PLoS One. 2014;9(6):e95034.
67. Lemey P, Kosakovsky Pond SL, Drummond AJ, Pybus OG, Shapiro B, Barroso H, et al. Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. PLoS Comput Biol. 2007;3(2):e29.
68. Amei A, Smith BT. Robust estimates of divergence times and selection with a poisson random field model: a case study of comparative phylogeographic data. Genetics. 2014;196(1):225–33.

69. Sawyer SA, Hartl DL. Population genetics of polymorphism and divergence. Genetics. 1992;132(4):1161–76.
70. Besenbacher S, Mailund T, Schierup MH. Association mapping and disease: evolutionary perspectives. Methods Mol Biol. 2012;856:275–91.
71. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248–9.
72. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 2008;4(5):e1000083.
73. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19(9):1553–61.
74. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747–53.
75. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, et al. Clinical assessment incorporating a personal genome. Lancet. 2010;375(9725):1525–35.
76. Stranger BE, Stahl EA, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics. 2011;187(2):367–83.
77. Phillips PC. Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet. 2008;9(11):855–67.
78. Daub JT, Hofer T, Cutivet E, Dupanloup I, Quintana-Murci L, Robinson-Rechavi M, et al. Evidence for polygenic adaptation to pathogens in the human genome. Mol Biol Evol. 2013;30(7):1544–58.
79. Galtier N. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol. 2001;18(5):866–73.
80. Khoury GA, Smadbeck J, Kieslich CA, Floudas CA. Protein folding and de novo protein design for biotechnological applications. Trends Biotechnol. 2014;32(2):99–109.
81. Stadler T, Kuhnert D, Bonhoeffer S, Drummond AJ. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proc Natl Acad Sci U S A. 2013;110(1):228–33.
82. Leventhal GE, Gunthard HF, Bonhoeffer S, Stadler T. Using an epidemiological model for phylogenetic inference reveals density dependence in HIV transmission. Mol Biol Evol. 2014;31(1):6–17.
83. Kuhnert D, Stadler T, Vaughan TG, Drummond AJ. Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death SIR model. J R Soc Interface/R Soc. 2014;11(94):20131106.
84. Campbell PJ, Pleasance ED, Stephens PJ, Dicks E, Rance R, Goodhead I, et al. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proc Natl Acad Sci U S A. 2008;105(35):13081–6.
85. Yates LR, Campbell PJ. Evolution of the cancer genome. Nat Rev Genet. 2012;13(11):795–806.
86. Litman GW, Cannon JP, Dishaw LJ. Reconstructing immune phylogeny: new perspectives. Nat Rev Immunol. 2005;5(11):866–79.
87. Zhu J, Ofek G, Yang Y, Zhang B, Louder MK, Lu G, et al. Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. Proc Natl Acad Sci U S A. 2013;110(16):6470–5.
88. Mirsky A, Kazandjian L, Anisimova M. Antibody-specific amino acid substitution model for bioinformatic inferences from antibody sequences. Mol Biol Evol. 2015;32(3):806–19.
89. Ma Q, Reeves JH, Liberles DA, Yu L, Chang Z, Zhao J, et al. A phylogenetic model for understanding the effect of gene duplication on cancer progression. Nucleic Acids Res. 2014;42(5):2870–8.
90. Schaper E, Gascuel O, Anisimova M. Deep conservation of human protein tandem repeats within the eukaryotes. Mol Biol Evol. 2014;31(5):1132–48.
91. Schaper E, Anisimova M. The evolution and function of protein tandem repeats in plants. New Phytol. 2014;206:397–410.
92. Hawkins NJ, Cools HJ, Sierotzki H, Shaw MW, Knogge W, Kelly SL, et al. Paralog re-emergence: a novel, historically contingent mechanism in the evolution of antimicrobial resistance. Mol Biol Evol. 2014;31(7):1793–802.
93. Luksza M, Lassig M. A predictive fitness model for influenza. Nature. 2014;507(7490):57–61.
94. Zoller S, Schneider A. Empirical analysis of the most relevant parameters of codon substitution models. J Mol Evol. 2010;70(6):605–12.
95. Stein LD. Bioinformatics: alive and kicking. Genome Biol. 2008;9(12):114.
96. Lee S, Jia Y, Jia M, Gealy DR, Olsen KM, Caicedo AL. Molecular evolution of the rice blast resistance gene Pi-ta in invasive weedy rice in the USA. PLoS One. 2011;6(10):e26260.
97. Joseph SJ, Didelot X, Rothschild J, de Vries HJ, Morre SA, Read TD, et al. Population genomics of Chlamydia trachomatis: insights on drift, selection, recombination, and population structure. Mol Biol Evol. 2012;29(12):3933–46.
98. Cornejo OE, Lefebure T, Bitar PD, Lang P, Richards VP, Eilertson K, et al. Evolutionary and population genomics of the cavity causing bacteria Streptococcus mutans. Mol Biol Evol. 2013;30(4):881–93.
99. Attard CR, Beheregaray LB, Jenner KC, Gill PC, Jenner MN, Morrice MG, et al. Hybridization of Southern Hemisphere blue whale subspecies and a sympatric area off Antarctica: impacts of whaling or climate change? Mol Ecol. 2012;21(23):5715–27.
100. Pauls SU, Nowak C, Balint M, Pfenninger M. The impact of global climate change on genetic diversity within populations and species. Mol Ecol. 2013;22(4):925–46.