BMC Bioinformatics

RESEARCH Open Access

Directional Darwinian Selection in proteins

David A McClellan[1,2]

From The 9th Annual Biotechnology and Bioinformatics Symposium (BIOT 2012)
Provo, UT, USA. 25-26 October 2012

Abstract

Background: Molecular evolution is a very active field of research, with several complementary approaches, including *dN/dS*, HON90, MM01, and others. Each has documented strengths and weaknesses, and no one approach provides a clear picture of how natural selection works at the molecular level. The purpose of this work is to present a simple new method that uses quantitative amino acid properties to identify and characterize *directional* selection in proteins.

Methods: Inferred amino acid replacements are viewed through the prism of a single physicochemical property to determine the amount and direction of change caused by each replacement. This allows the calculation of the probability that the mean change in the single property associated with the amino acid replacements is equal to zero (H_0: $\mu = 0$; i.e., no net change) using a simple two-tailed t-test.

Results: Example data from calanoid and cyclopoid copepod cytochrome oxidase subunit I sequence pairs are presented to demonstrate how directional selection may be linked to major shifts in adaptive zones, and that convergent evolution at the whole organism level may be the result of convergent protein adaptations.

Conclusions: Rather than replace previous methods, this new method further complements existing methods to provide a holistic glimpse of how natural selection shapes protein structure and function over evolutionary time.

Background

Natural selection, as first outlined by Charles Darwin, acts on phenotypes:

"She [natural selection] can act on every internal organ, on every shade of constitutional difference, on the whole machinery of life...Under nature, the slightest difference of structure or constitution may well turn the nicely-balanced scale in the struggle for life, and so be preserved...It may be said that natural selection is daily and hourly scrutinizing, throughout the world, every variation, even the slightest; rejecting that which is bad, preserving and adding up all that is good." [1].

We can think of natural selection as collecting adaptations that optimize an organism's survival, reproductive success, and fecundity in a given environment or habitat. As Darwin explicitly states above, this process is not limited to the phenotypes of the whole organism; it works on "every variation, even the slightest." Although we sometimes think of proteins in this way, there currently is not a consistently reliable method for identifying and characterizing the evolution of protein phenotypes. This being stated, science is currently faced with the challenge of assessing the impact that anthropogenic climate change is likely to have with potentially catastrophic effects at the base of the food chain on the molecular level. The scientific community's efforts to produce realistic solutions to the big problems associated with climate change will be greatly enhanced by the development of more robust analytical methods for comprehensively characterizing the effects of natural selection in terms of the biochemistry and physics of protein structure, function, and interactions.

Several statistical methods for identifying and characterizing selection at the molecular level have been proposed since the genetic code was determined in the 1960s. Of these, three classes of methods dominate the literature. The first, and most significant, is the family of methods that implements one of many variations of the nonsynonymous-to-synonymous substitution rate ratio,

Correspondence: david.mcclellan@uafs.edu
[1] Department of Biology, University of Arkansas-Fort Smith, Fort Smith, AR 72913, USA
Full list of author information is available at the end of the article

or *dN*/*dS* (e.g., [2-15]). Briefly, this approach compares the rate of nonsynonymous (*dN*), or amino acid changing, nucleotide substitutions with the rate of synonymous (*dS*), or silent, nucleotide substitution. When *dN* is significantly greater than *dS*, the system is said to have been influenced by positive selection, when *dN* = *dS*, the system is said to be neutral, and *dN* <*dS* indicates negative selection. This family of methods is broadly accepted and implemented, and has enjoyed a great deal of success. This simple model, however, has several shortcomings, including problems with the underlying assumptions (e.g., [16-19]) and difficulties accurately estimating rates when divergences are very small and great, and is not sensitive enough to detect natural selection in some protein coding genes when it is known to have taken place (e.g., [20-22]).

As a reaction to weaknesses of *dN*/*dS* approaches, Hughes et al. [23] presented a similar approach (hereafter referred to as HON90) that compares proportions of conservative (p_{NC}) and radical (p_{NR}) amino acid replacement in terms of *qualitative* properties of amino acids to detect selection promoting charge profile diversity in class I MHC proteins. When $p_{NR} > p_{NC}$, the property of interest is said to have changed more than would be expected under random conditions. The Hughes et al. study [23] was the first to implement amino acid properties–in this case charge, polarity, hydrophobicity, and volume–to identify selection at the protein level. From a conceptual standpoint, this approach presented a method to assess patterns of amino acid replacement using the phenotypes of proteins, thus providing an avenue of analysis more consistent with Darwin's original definition of natural selection. This protein-level phenotypic approach has since been implemented several times and has yielded encouraging results (e.g., [24-28]).

In an effort to tap into the wealth of information afforded by the implementation of *quantitative* amino acid properties, researchers have expanded upon the HON90 approach in a number of ways, including the use of a spectrum of magnitude categories [18,29], a sliding window [30], accuracy benchmarking [31], and potential uses for characterizing single amino acid replacements [32]. These approaches (hereafter referred to as MM01 methods) take the underlying pattern of nucleotide composition into account. The collective identity of properties that individually yield positive statistical results provides clues that link specific genetic variants to selective advantages and disadvantages afforded by known changes in ambient environment [18,33,34]. The robustness of the results yielded by MM01 approaches is greatly enhanced by the wealth of information emerging from crystallography and magnetic resonance experiments that determine protein structures with a high degree of precision and accuracy. Results localized to protein regions of known structures and functions provide evidence useful for comprehensively characterizing protein function evolution [30,34-38].

Existing solutions fall short

Reconstructing evolutionary events at the molecular level and diagnosing them in terms of natural selection has been an extraordinary challenge. Each individual point mutation carries with it just a small quantum of information. Patterns emerge as these quanta accumulate over evolutionary time. Oversimplifying models used to assess patterns of evolutionary information emerging from molecular data results in a net loss of analytical yield. Overparameterizing models has the opposite effect, producing more detail than can be realistically supported by the data. When studying a phenomenon as nuanced and multifaceted as molecular evolution, striking a happy medium between oversimplification and overparameterization is extremely difficult. Researchers want to squeeze every ounce of information from their data without seeing patterns that are not really there.

It is not surprising that *dN*/*dS* approaches sometimes ignore signs of natural selection that other methods pick up. *dN*/*dS* is a simple method, with several documented limitations [16-22]. The HON90 approach takes a step forward by incorporating amino acid properties, but the number of qualitative properties is limited; if the evolution of protein-coding gene sequences cannot be linked to charge, polarity, hydrophobicity, volume, or just a handful of other properties, negative results will be produced. Although these properties are important in terms of protein function, they likely are not the only properties affected by natural selection.

MM01 approaches present several advantages over *dN*/*dS* and HON90 methods (e.g., [18,30,35,36,39-45]). However, in an effort to force a greater information yield from the data, this method may be parameterizing systems to the point that accuracy suffers [31,32,46]. Clearly, this third class of approaches performs better in some circumstances, such as when divergences are very great and rates of synonymous change are underestimated [18], or when divergences are very small and synonymous changes have not had time to accumulate [32].

Methods

The high frequency with which new genomes and metagenomes are being produced also suggests that a method with the potential for high-throughput that does not require information from underlying nucleotides is needed. Gene annotations produce a huge number of BLAST results [47,48]. Many of these are in the form of aligned protein, and not nucleotide, sequences. None of the methods outlined above are capable of screening this type of information for signs of molecular adaptation and cannot be utilized for studying adaptive changes at the genomic or metagenomic levels.

There is at least one aspect of physicochemical evolution that has been largely overlooked: *the direction of selection*. One exception is the study by Merritt and Quattro [27]. They identified a case in which positive selection resulted in a biased accumulation of negatively charged amino acids after a gene duplication event. However, changes in charge are generally rare in protein evolution [27,49,50] and, as discussed, the possible qualitative properties to test in the way Merritt and Quattro present are few in number. Testing for directional shifts in *quantitative* properties, of which there are now several hundred catalogued in the Japanese database AAindex [51], will allow for more comprehensive exploration of property space, and will likely result in a more clearly resolved vision of how proteins adapt to the specific needs of organisms as they evolve in changing habitats. Such a new method, when coupled with existing methods, will provide a full set of analytical tools for identifying and characterizing molecular adaptation in a biologically meaningful way.

A method similar to that presented by Merritt and Quattro [27] that allows for the implementation of quantitative physicochemical amino acid properties will require a different statistical test. Inferred amino acid replacements will be viewed through the prism of a single physicochemical property to determine the amount and direction of change caused by each replacement. This will allow the calculation of the probability that the mean change in the single property associated with the amino acid replacements is equal to zero (H_0: $\mu = 0$; i.e., no net change) using a simple two-tailed t-test.

The novel aspect of this new method is its criterion. It evaluates amino acid replacements multi-dimensionally across a great number of physicochemical amino acid properties, and identifies instances of several amino acid replacements across several sites, evolving across phylogenetic space in the same physicochemical direction in a single dimension of property space. This approach makes the study of molecular evolution more applicable to studies that link patterns of amino acid replacement with environmental changes through time or space. A directional approach represents a return to the fundamental concept that selection affects phenotypes, while at the same time simplifying implementation. By so doing, interpretation of results will be less ambiguous.

The new method begins with a list of amino acid differences that includes the location of each in the context of the linear sequence of nucleotide codons and/or amino acids, depending on the input data. This list can be generated using an ancestral character-state reconstruction algorithm (such as *codeml* [52]) if the input is a multiple sequence alignment and a phylogenetic structure, or by pairwise comparison if the input is the results of a BLAST search [47,48]. From this list, the magnitude and direction (i.e., an increase or a decrease) of change in each amino acid property under consideration is inferred. A simple two-tailed t-test may be performed for each property to statistically evaluate the null hypothesis that the net change is equal to zero. The value of the t-test statistic is calculated using simple established equations:

$$t = \frac{\bar{X}}{s_{\bar{X}} \big/ \sqrt{N}} \tag{1}$$

$$s_{\bar{X}} = \sqrt{\frac{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{N}}{N-1}} \tag{2}$$

Here X_i is the value of the change in amino acid property for each inferred amino acid difference, i, and N is the total number of amino acid differences. In the example below (Table 1), the value of X_i for the difference at residue site 82 is +7.0, while the value of N is 15.

The data may be partitioned in several different ways: A sliding window may be implemented to evaluate potential clustering of unidirectional changes; known or estimated secondary structures may be used to group amino acid differences according to the structural components of the protein; the range of amino acid sites corresponding to the functional domains of the protein may be used. How the data are partitioned is largely contingent on the scientific question, the amount and type of differences in the data, and the amount of supporting structure and function

Table 1 Directional selection analysis of *Pan* and *Homo* SAGE1

Residue	*Pan*	*Homo*	Δ Hydropathy
82	Cys	Arg	+7.0
92	Val	Ala	+2.4
160	Arg	His	-1.3
450	Gln	Arg	+1.0
507	Asp	Val	-7.7
523	Ser	Thr	-0.1
563	Val	Ala	+2.4
582	Val	Asp	+7.7
605	Phe	Leu	-1.0
672	Ala	Thr	+2.5
675	Ser	Asn	+2.7
694	Thr	Ala	-2.5
754	Cys	Arg	+7.0
802	Val	Ala	+2.4
805	Leu	Ser	+4.6
		Net Change =	+27.1

Residue sites of the 15 *Pan troglodytes* and *Homo sapiens* SAGE1 (inferred from XM_001137139 and NM_018666, respectively) protein differences, with character states and net change in hydropathy [54].

information available. In each case, care must be taken to partition the data in biologically meaningful ways that test specific hypotheses.

There are over 500 physicochemical amino acid properties on the AAindex database [51] available to assess amino acid differences. For the purposes of this study, the 25 properties in Table 2 were chosen to be representative of the breadth of amino acid property space. These properties describe aspects of proteins that are important to overall structure (e.g., molecular size, hydrophobicity, secondary structures) and function (e.g., ionization, non-bonded energy, solvent accessibility); properties that can potentially be affected by natural selection.

Together, these four complementary methods will enable more robust evaluation of data than is possible with any single method: *dN*/*dS* methods focus on patterns of nucleotide substitution; HON90 looks at phenotypic patterns across amino acid changes; MM01 methods emphasize patterns among the most radical changes; the new method detects localized directional shifts in protein phenotypes. Furthermore, certain methods are able to more easily accommodate different data types. All of the methods can assess multiple protein-coding nucleotide sequence alignments with an accompanying phylogenetic structure, but *dN*/*dS* methods, for example, are unable to evaluate *blastp* output because there is no way to estimate the rate of synonymous change in the encoding DNA sequences from aligned amino acid sequences. The new directional selection method will easily accept *blastp* output because it does not require information about the underlying pattern of nucleotides or the governing genetic code.

Results and discussion

Directional selection linked to Habitat Shifts

Several marine and freshwater calanoid copepod cytochrome oxidase subunit I (COI) sequence pairs were considered. The first approximately 650 nucleotides of the cytochrome oxidase subunit 1 coding region for each were obtained from the Barcode of Life Database (http://www.barcodinglife.com) and evaluated using the directional selection approach. The comparison of *Calanus hyperboreus* (marine) and *Mastigodiaptomus montezumae* (freshwater) is representative (GenBank accession numbers FJ602504 and EU770508, respectively). Interestingly, the first 650 nucleotides encode all of the components of the first COI proton pump [53]. There are 11 amino acid differences within the first 215 amino acid residue sites for this species pair. These replacements have resulted in radical changes in several physicochemical properties. None of the properties were implicated in the proton output region of the protein ($p < 0.05$), but three properties affected the proton input region: one that describes hydrophobicity (H_p), one for polarity (P_r), and one for tertiary structure (F). Collectively, these properties, coupled with their direction of change, indicate that the proton input region became less hydrophobic, more polar, and more structural malleability during calanoid adaptation to freshwater, resulting in a more direct and less energetically expensive path for hydrogen ions to penetrate the membrane and enter the proton pump.

Several marine and freshwater cyclopoid copepod cytochrome oxidase subunit I (COI) sequence pairs were considered as well. Of these, the comparison of *Oithona similis* (marine) and *Thermocyclops inversus* (freshwater) is representative (GenBank accession numbers EU599544 and EU770551, respectively). There are 40 amino acid differences within the first 215 residue sites of COI for this species pair. Five properties yielded statistically significant directional results ($p < 0.05$) across the entire alignment, including V^0, P_r, p, μ, and H_t. Like the calanoid data, the cyclopoid data failed to exhibit positive results in the proton output region. The proton input region, however, experienced significant directional change in 12 properties (Table 3). The identity of the properties and the direction of change were similar to the calanoid results, indicating a decrease in hydrophobicity (h, H_p, H_t), an increase in polarity (P_r, p), and increased structural malleability (N_a, B_r, F), but cyclopoids also exhibited a decrease in molecular size (B_l, V^0) and total non-bonded energy (E_t), and an increase in turn tendency (P_t). Collectively, these results suggest an even more direct and less energetically expensive path for hydrogen ions to enter the proton pump than exhibited by the calanoids.

Interestingly, the calanoid and cyclopoid results appear parallel at the property level even though none of the specific sites affected were the same. To illustrate even the subtle parallel shifts in properties, Table 3 also includes those properties that yielded results at a lower significance ($p = 0.1$). Every property affected during calanoid adaptation to freshwater was also affected during cyclopoid adaptation to freshwater, and in the same direction. Cyclopoids had a greater number of affected properties likely due to a greater accumulation of amino acid replacements.

The discovery that these two lineages of copepods found parallel routes for COI functional adaptation is the most exciting conclusion of these results. These findings suggest that the amazing amount of convergence in the natural world may be the result of a limited number of alternative physicochemical strategies. This partially explains how independently evolving proteins can converge upon similar structures and functions when sequence identity remains low. Furthermore, the consistency of these results

Table 2 Quantitative physicochemical amino acid properties

Category/Property	Symbol	Reference	Description[a]
Hydrophobicity			
Hydropathy	h	[54]	The hydrophilic and hydrophobic inclinations of a given residue side chain in terms of transfer of free energy.
Surrounding hydrophobicity	H_p	[55]	The average sum of residue hydrophobic indices within an optimum sphere of 8 Å radius around a residue in protein crystals (kcal/mol).
Thermodynamic transfer hydrophobicity	H_t	[56]	The experimental values of Noazaki & Tanford [57] combined with values of Zimmerman et al. [58] adjusted to the same scale (kcal/mol).
Ionization Constants			
Equilibrium constant	pK'	[56]	The ionizable character of the carboxyl group (pH units).
Isoelectric point	pH_i	[58]	The isoionic point of the free amino acid, including the ionizable character of the entire residue (pH units).
Molecular Size & Composition			
Bulkiness	B_l	[58]	The ratio of the side chain volume to length (i.e., the average cross section of the chain) (Å^2).
Composition	c	[59]	The atomic weight ratio of the non-carbon elements in the end groups or rings to carbons in the side chain.
Molecular weight	M_w	[60]	The mass of the atoms constituting the residue.
Partial specific volume	V^0	[61]	The reciprocal of density (m^3 mol^{-1} × 10^{-6}).
Non-bonded Energy			
Long range energy	E_l	[62]	The energy between two amino acids separated further than 10 residues (i.e., due to electrostatic and Van der Waals forces) (kcal/mol).
Short and medium range energy	E_{sm}	[62]	The sum of the energy between 1) main chain atoms of a residue and its own side chain atoms, and 2) two residues located within 10 residues along the chain (kcal/mol).
Total non-bonded energy	E_t	[62]	Sum of average short, medium, and long range non-bonded energies (kcal/mol).
Polarity & Polarizability			
Polar requirement	P_r	[63]	The slope of the line regressing log R_M and the mol fraction of water in the pyrimidine-water solvent (R_M = 1/R_F - 1, where R_F is the chromatographic index [58]).
Polarity	p	[59]	The average of P_r and P_A (P_A = 13.66 - 14.85R_F).
Refractive index	μ	[55]	The measure of the polarizability of a residue (i.e., the reciprocal measure of its electrical stability under an external field).
Secondary Structure			
Alpha-helical tendency	P_α	[64]	The average intrinsic property of a residue to adopt an alpha-helical conformation.
Beta-structure tendency	P_β	[64]	The average intrinsic property of a residue to adopt a beta-sheet conformation.
Coil tendency	P_c	[65]	A measure of the tendency that a particular residue will be found in a coil secondary structure.
Helical contact area	C_a	[66]	The maximum area loss that could occur in going from an isolated α-helix to a fully buried environment in the complex (Å^2).
Turn tendency	P_t	[64]	The average intrinsic property of a residue to adopt a beta-turn conformation.
Solvent Accessibility			
Solvent accessibility reduction ratio	R_a	[67]	The ratio of solvent accessibility: the solvent accessibility of a residue in a hypothetically extanded state over its accessibility in a native folded protein.
Tertiary Structure			
Average number of surrounding residues	N_a	[67]	The average number of residues surrounding a residue within the effective distance of influence.
Buriedness	B_r	[61]	The average propensity of a residue to be found in the interior of a protein.
Compressibility	K^0	[61]	The relative increase in the volume of the system per unit decrease in pressure (m^3 mol^{-1} P_a^{-1} × 10^{-15}).
Mean rms fluctuational displacement	F	[68]	The relationship between the average amount of root-mean-square displacement of a residue and its distance from the centroid of the protein (Å).

[a] Properties without defined units are dimensionless.

Twenty-five amino acid properties representative of the breadth of amino acid property space.

demonstrates how analyzing protein-coding genes in terms of changing protein phenotypes provides a link between the evolution of organisms and the influence of environmental variables, and hints at the actual causes of natural selection.

Conclusions

The methods for identifying and characterizing natural selection at the molecular level, dN/dS, HON90, and MM01, use different aspects of the evolutionary information locked in protein-coding sequencing sequences.

Table 3 Results of directional selection analysis of marine and freshwater copepod COI

	Calanoids		Cyclopoids	
Properties	**Proton Input**	**Proton Output**	**Proton Input**	**Proton Output**
Hydrophobicity				
h	- - -			
H_p	- -		- -	
H_t	-		-	
Ionization Constants				
pK'				
pH_i				
Molecular Size				
B_l	- - -			
c				
M_w	-			
V^0	- - -		-	
Non-bonded Energy				
E_l	-			
E_{sm}				
E_t	- -			
Polarity & Polarizability				
P_r	+ + +		+ +	
p	+ + +		+	
μ	- -		-	
Secondary Structure				
P_α				
P_β				
P_c	+			
C_a	-		-	
P_t	+ +			
Solvent Accessibility				
R_a				
Tertiary Structure				
N_a	- - -			
B_r	- -		-	
K^0	+			
F	+ + +		+ +	

Properties among the 25 defined in Table 2 that experienced significant directional shifts as copepods evolved from being adapted to a marine habitat to life in freshwater ("+" indicates significant increases in the property, while "-" indicates significant decreases in the property; one symbol indicates significance at the α = 0.1 level, while two symbols indicate significance at the α = 0.05 level and three indicates α = 0.02 significance).

However, none of these methods are able to identify signs of adaptation in protein sequences without the aid of the underlying nucleotide information. A new method for identifying adaptation in either protein *or* protein-coding DNA sequences is presented. Collectively, the four methods will enable a more robust evaluation of existing data than is possible with any single method. Furthermore, the new directional selection method can tap the wealth of information in BLAST reports, like those emerging from genome and metagenome annotation efforts. It is likely that high-throughput analysis of annotation reports will provide a glimpse of the collective evolutionary forces that shape the morphologies and behaviors at the organismal level, especially as they evolve in a changing environment, providing a strong link between macroevolution and microevolution. Such a link will likely prove important to improving our understanding of how modern anthropogenic changes in global and local climates may be affecting vulnerable organisms over evolutionary time or at more accelerated rates.

Competing interests

There are no competing interests with regard to this research.

Authors' contributions

This research was performed by the sole author.

Acknowledgements

The author thanks Bigelow Laboratory for Ocean Sciences for support of this research. He also a acknowledges the many students who have worked on associated projects, especially S Woolley, J K Sailsbery, R G Christensen, A Fuchsman, and A Burchill. Furthermore, he thanks M J Clement and K A Crandall for numerous conversations and assistance mentoring students. He thanks G Wyngaard for help understanding copepods and their gene sequences. Finally, he thanks G Shimmield and T Brailovskaya for advice and editing assistance.

Declarations

Funding for publication was provided by the Department of Biological Science at University of Arkansas-Fort Smith.

This article has been published as part of BMC Bioinformatics Volume 14 Supplement 13, 2013: Selected articles from the 9th Annual Biotechnology and Bioinformatics Symposium (BIOT 2012). The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements/14/S13

Authors' details

[1]Department of Biology, University of Arkansas-Fort Smith, Fort Smith, AR 72913, USA. [2]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME 04544, USA.

Published: 1 October 2013

References

1. Darwin C: *On the Origin of Species* Harvard University Press, Cambridge, Massachusetts; 1964, 83-84, (facsimile of the First Edition, 1859).
2. Nei M, Gojobori T: **Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.** *Molecular Biology and Evolution* 1986, **3**:418-426.
3. Lee Y-H, Vacquier VD: **The divergence of species-specific abalone sperm lysins is promoted by positive Darwinian selection.** *Biological Bulletin* 1992, **182**:97-104.
4. Li W-H: **Unbiased estimation of the rates of synonymous and nonsynonymous substitution.** *J Molecular Evolution* 1993, **36**:96-99.
5. Nielsen R, Yang Z: **Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene.** *Genetics* 1998, **148**:929-936.
6. Yang Z, Nielsen R, Goldman N, Pedersen A-MK: **Codon-substitution models for heterogeneous selection pressure at amino acid sites.** *Genetics* 2000, **155**:431-449.
7. Swanson WJ, Yang Z, Wolfner MF, Aquadro CF: **Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals.** *Proceedings of the National Academy of Sciences, USA* 2001, **98**:2509-2514.
8. Fares MA, Wolfe KH: **Positive selection and subfunctionalization of duplicated CCT chaperonin subunits.** *Molecular Biology and Evolution* 2003, **20**:1588-1597.

9. Chen L, Perlina A, Lee CJ: **Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase.** *J Virology* 2004, **78**:3722-3732.
10. Filip LC, Mundy NI: **Rapid evolution by positive Darwinian selection in the extracellular domain of the abundant lymphocyte protein CD45 in primates.** *Molecular Biology and Evolution* 2004, **21**:1504-1511.
11. Pogson GH, Mesa KA: **Positive Darwinian selection at the pantophysin (*Pan* I) locus in marine gadid fishes.** *Molecular Biology and Evolution* 2004, **21**:65-75.
12. Petersen L, Bollback JP, Dimmic M, Hubisz M, Nielsen R: **Genes under positive selection in *Escherichia coli*.** *Genome Research* 2007, **17**:1336-1343.
13. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES, The International HapMap Consortium: **Genome-wide detection and characterization of positive selection in human populations.** *Nature* 2007, **449**:913-918.
14. Kosiol C, Vinař T, de Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A: **Patterns of positive selection in six mammalian genomes.** *PLoS Genetics* 2008, **4**:e1000144.
15. Metzger KJ, Thomas MA: **Evidence of positive selection at codon sites localized in extracellular domains of mammalian CC motif chemokine receptor proteins.** *BMC Evolutionary Biology* 2010, **10**:139.
16. Hughes AL: *Adaptive Evolution of Genes and Genomes* Oxford University Press, Oxford, UK; 1999.
17. Gillespie JH: *Population Genetics: A Concise Guide* Johns Hopkins University Press, Baltimore, Maryland; 2004.
18. McClellan DA, Palfreyman EJ, Smith MJ, Moss JL, Christensen RG, Sailsbery JK: **Physicochemical evolution and molecular adaptation of the cetacean and artiodactyl cytochrome *b* proteins.** *Molecular Biology and Evolution* 2005, **22**:437-455.
19. Hughes AL: **Looking for Darwin in all the wrong places: The misguided quest for positive selection at the nucleotide sequence level.** *Heredity* 2007, **99**:364-373.
20. Wolfe KH, Sharp PM: **Mammalian gene evolution: Nucleotide sequence divergence between mouse and rat.** *Journal of Molecular Evolution* 1993, **37**:441-456.
21. Crandall KA, Kelsey CR, Imamichi H, Lane HC, Salzman NP: **Parallel evolution of drug resistance in HIV: Failure of nonsynonymous/ synonymous substitution rate ratio to detect selection.** *Molecular Biology and Evolution* 1999, **16**:372-382.
22. Creevey CJ, McInerney JO: **An algorithm for detecting directional and non-directional positive selection, neutrality and negative selection in protein coding DNA sequences.** *Gene* 2002, **300**:43-51.
23. Hughes AL, Ota T, Nei M: **Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of Class I major-histocompatibility-complex molecules.** *Molecular Biology and Evolution* 1990, **7**:515-524.
24. Swanson WJ, Vacquier VD: **Extraordinary divergence and positive Darwinian selection in a fusagenic protein coating the acrosomal process of abalone spermatozoa.** *Proceedings of the National Academy of Sciences, USA* 1995, **92**:4957-4961.
25. Metz EC, Palumbi SR: **Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein bindin.** *Molecular Biology and Evolution* 1996, **13**:397-406.
26. Zhang J: **Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes.** *Journal of Molecular Evolution* 2000, **50**:56-68.
27. Merritt TJS, Quattro JM: **Evidence for a period of directional selection following gene duplication in a neutrally expressed locus of triosephosphate isomerase.** *Genetics* 2001, **159**:689-697.
28. Van de Peer Y, Taylor JS, Braasch I, Meyer A: **The ghost of selection past: Rates of evolution and functional divergence of anciently duplicated genes.** *Journal of Molecular Evolution* 2001, **53**:436-446.
29. McClellan DA, McCracken KG: **Estimating the influence of selection on the variable amino acid sites of the cytochrome *b* protein functional domains.** *Molecular Biology and Evolution* 2001, **18**:917-925.
30. Porter ML, Cronin TW, McClellan DA, Crandall KA: **Molecular characterization of crustacean visual pigments and the evolution of pancrustacean opsins.** *Molecular Biology and Evolution* 2007, **24**:253-268.
31. McClellan DA, Ellison DD: **Assessing and improving the accuracy of detecting protein adaptation with the TreeSAAP analytical software.** *International J Bioinformatics Research and Application* 2010, **6**:120-133.
32. McClellan DA: **Detecting molecular selection on single amino acid replacements.** *International J Bioinformatics Research and Applications* 2012, **8**:67-80.
33. Chamala S, Beckstead WA, Rowe MJ, McClellan DA: **Evolutionary selective pressure on three mitochondrial SNPs is consistent with their influence on metabolic efficiency in Pima Indians.** *International J Bioinformatics Research and Applications* 2007, **3**:504-522.
34. Beckstead WA, Ebbert MTW, Rowe MJ, McClellan DA: **Evolutionary pressure on mitochondrial cytochrome *b* is consistent with a role of cytbI7T affecting longevity during caloric restriction.** *PLoS ONE* 2009, **4**:e5836.
35. da Fonseca RR, Antunes A, Melo A, Ramos MJ: **Structural divergence and adaptive evolution in mammalian cytochromes P450 2C.** *Gene* 2007, **387**:58-66.
36. Osorio DS, Antunes A, Ramos MJ: **Structural and functional implications of positive selection at the primate angiogenin gene.** *BMC Evolutionary Biology* 2007, **7**:167.
37. Castoe TA, Jiang ZJ, Gu W, Wang ZO, Pollock DD: **Adaptive evolution and functional redesign of core metabolic proteins in snakes.** *PLoS ONE* 2008, **3**:e2201.
38. da Fonseca RR, Johnson WE, O'Brien SJ, Ramos MJ, Antunes A: **The adaptive evolution of the mammalian mitochondrial genome.** *BMC Genomics* 2008, **9**:119.
39. Pérez-Losada M, Viscidi RP, Demma JC, Zenilman J, Crandall KA: **Population genetics of *Neisseria gonorrhoeae* in a highprevelence community using a hypervariable outer membrane *porB* and 13 slowly evolving housekeeping genes.** *Molecular Biology and Evolution* 2005, **22**:1887-1902.
40. Pérez-Losada M, Browne EB, Madsen A, Wirth T, Viscidi RP, Crandall KA: **Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data.** *Infection, Genetics and Evolution* 2006, **6**:97-112.
41. Pérez-Losada M, Crandall KA, Bash MC, Dan M, Zenilman J, Viscidi RP: **Distinguishing importation from diversification of quinolone-resistant *Neisseria gonorrhoeae* by molecular evolutionary analysis.** *BMC Evolutionary Biology* 2007, **7**:84.
42. Taylor SD, Dittmar de la Cruz K, Porter ML, Whiting MF: **Characterization of the long-wavelength opsin from Mecoptera and Siphonaptera: Does a flea see?** *Molecular Biology and Evolution* 2005, **22**:1165-1174.
43. Marques AT, Antunes A, Fernandes PA, Ramos MJ: **Comparative evolutionary genomics of the HADH2 gene encoding Aβ-binding alcohol dehydrogenase/17β-hydroxysteroid dehydrogenase type 10 (ABAD/HSD10).** *BMC Genomics* 2006, **7**:202.
44. Streisfeld MA, Rausher MD: **Relaxed constraint and evolutionary rate variation between basic helix-loop-helix floral anthocyanin regulators in *Ipomoea*.** *Molecular Biology and Evolution* 2007, **24**:2816-2826.
45. Chapman EG, Piontkivska H, Walker JM, Stewart DT, Curole JP, Hoeh WR: **Extreme primary and secondary protein structure variability in the chimeric male-transmitted cytochrome c oxidase subunit II protein in freshwater mussels: Evidence for an elevated amino acid substitution rate in the face of domain-specific purifying selection.** *BMC Evolutionary Biology* 2008, **8**:165.
46. Maxwell TJ, Bendall ML, Staples J, Jarvis T, Crandall KA: **Phylogenetics applied to genotype/phenotype association and selection data from Angptl4 in humans.** *International J Molecular Sciences* 2010, **11**:370-385.
47. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic Local Alignment Search Tool.** *Journal of Molecular Biology* 1990, **215**:403-410.
48. Altschul SF, Boguski MS, Gish W, Wootton JC: **Issues in searching molecular sequence databases.** *Nature Genetics* 1994, **6**:119-129.
49. Peetz EW, Thomson G, Hedrick PW: **Charge changes in protein evolution.** *Molecular Biology and Evolution* 1986, **3**:84-94.
50. Xia X, Li W-H: **What amino acid properties affect protein evolution?** *J Molecular Evolution* 1998, **47**:557-564.
51. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M: **AAindex: Amino acid database progress report 2008.** *Nucleic Acids Research* 2008, **36**:D202-D205.
52. Yang Z: **PAML 4: Phylogenetic Analysis by Maximum Likelihood.** *Molecular Biology and Evolution* 2007, **24**:1586-1591.
53. Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi H, Shinzawa-Itoh K, Nakashima R, Yaono R, Yoshikawa S: **The whole structure of the 13-subunit oxidized cytochrome *c* oxidase at 2.8 Å.** *Science* 1996, **272**:1136-1144.
54. Kyte J, Doolittle RF: **A simple method for displaying the hydropathic character of a protein.** *J Molecular Biology* 1982, **157**:105-132.

55. Probhakaran M, Ponnuswamy PK: **The spatial distribution of physical, chemical, energetic and conformational properties of amino acid residues in globular proteins.** *J Theoretical Biology* 1979, **80**:485-504.
56. Jones DD: **Amino acid properties and side-chain orientation in proteins: A cross correlation approach.** *J Theoretical Biology* 1975, **50**:167-183.
57. Nozaki Y, Tanford C: **The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions: Establishment of a hydrophobicity scale.** *Journal of Biological Chemistry* 1971, **246**:2211-2217.
58. Zimmerman JM, Eliezer N, Simha R: **The characterization of amino acid sequences in proteins by statistical methods.** *J Theoretical Biology* 1968, **21**:170-201.
59. Grantham R: **Amino acid difference formula to help explain protein evolution.** *Science* 1974, **185**:862-864.
60. Fasman GD: *Handbook of Biochemistry and Molecular Biology.* 3 edition. CRC Press, Cleveland, Ohio; 1976, Proteins - Volume 1.
61. Gromiha MM, Ponnuswamy PK: **Relationship between amino acid properties and protein compressibility.** *J Theoretical Biology* 1993, **165**:87-100.
62. Oobatake M, Ooi T: **An analysis of non-bonded energy of proteins.** *J Theoretical Biology* 1977, **67**:567-584.
63. Woese CR: **Evolution of the genetic code.** *Naturwissenschaften* 1973, **60**:447-459.
64. Chou PY, Fasman GD: **Prediction of the secondary structure of proteins from their amino acid sequence.** *Advances in Enzymology and Related Areas of Molecular Biology* 1978, **47**:45-148.
65. Charton M, Charton B: **The dependence of the Chou-Fasman parameters on amino acid side chain structure.** *J Theoretical Biology* 1983, **102**:121-134.
66. Richmond TJ, Richards FM: **Packing of α-helices: Geometrical constraints and contact areas.** *J Molecular Biology* 1978, **119**:537-555.
67. Ponnuswamy PK, Prabhakaran M, Manavalan P: **Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins.** *Biochimica et Biophysica Acta* 1980, **623**:301-316.
68. Bhaskaran R, Ponnuswamy PK: **Dynamics of amino acid residues in globular proteins.** *International J Peptide and Protein Research* 1984, **24**:180-191.

doi:10.1186/1471-2105-14-S13-S6
Cite this article as: McClellan: Directional Darwinian Selection in proteins. *BMC Bioinformatics* 2013 **14**(Suppl 13):S6.