J. theor. Biol. (1997) **188,** 79–87

Cultural Evolution and Units of Selection in Replicating Text

Richard Pocklington* and Michael L. Best†‡

* *Department of Anthropology, Human Biocultural Evolution Research Group, Stanford University, Stanford, CA* 94305, *U.S.A.*; *and the* † *Media Laboratory, Massachusetts Institute of Technology, E*15-320*F, Cambridge, MA* 02139, *U.S.A.*

(*Received on* 18 *October* 1996, *Accepted in revised form on* 12 *April* 1997)

The use of biological models and metaphors in studies of culture has a long and checkered history. While there are many superficial similarities between biological and cultural evolution, attempts to pin down such analogies have not been wholly successful. One limiting factor may be a lack of empirical evidence that the basic assumptions of the evolutionary model are met within a cultural system. We argue that a focus on the detection and description of the units of selection is an essential first step in constructing any evolutionary model. In this paper we outline the necessary connection between units of selection and evolution, describe the properties of a unit of selection, and introduce an empirical method for the detection of putative units of selection in a model cultural system: discourse within NetNews, a discussion system on the Internet.

1. Introduction

> "One of the most interesting things about Darwin's explanation of the origin of species is that scarcely anything need be assumed about the actual nature of species, as evidence that natural selection occurs; the same process is in progress with respect to languages, religions, habits, customs, rocks, beliefs, chemical elements, nations, and everything else to which the terms stable and unstable can be applied. The only things required of a species are the capacities of variation and inheritance." R. A. Fisher (1912).

Richard Dawkins (1976, 1982) has convincingly argued what evolutionary biologists have claimed since the inception of the field: the theory of natural selection is not limited to biological systems (Darwin, 1859). Change due to selection is not a property of a particular mode of inheritance; it is a universal principle of self-replicating systems (Holland, 1975; Schuster & Sigmund, 1983). In this paper we apply the theory of natural selection to the differential propagation of cultural elements through a human social network. While many authors have investigated the similarities between human culture and self-replicating biological systems [e.g. Cloak, 1973; Plotkin, 1994; reviewed in Pocklington (1996)] few have taken the steps to construct an explicit model of cultural transmission (Cavalli-Sforza & Feldman, 1981; Lumsden & Wilson, 1981; Boyd & Richerson, 1985), and fewer still have attempted to test these models (Cavalli-Sforza *et al*., 1982; Hewlett & Cavalli-Sforza, 1986; Lynch *et al*., 1989).

Before we attempt to model cultural evolution, we must be clear about certain fundamental assumptions. While natural selection is a general process, not limited to biological systems, it is not a magical force to be invoked by fiat.

In this paper we first provide a brief description of the necessary conditions for natural selection to occur. Next we outline the connection between units of selection and evolution. We continue with a description of the necessary properties of a unit of selection. We then describe a statistical model based on text analysis which detects replicating patterns within a corpus. We conclude with some preliminary results from the use of this text analysis system.

‡ Author to whom correspondence should be addressed. E-mail: mikeb@media.mit.edu.

0022–5193/97/170079 + 09 $25.00/0/jt970460

MINIMAL PROPERTIES NEEDED FOR A SYSTEM TO EXHIBIT NATURAL SELECTION

Any system which has the properties of imperfect replication and trait/fitness covariance will be expected to undergo change due to selection (Lewontin, 1970); that is, change due to selection will occur within any non-homogenous population of replicating entities if they replicate with high fidelity yet still with some imperfections, and they transmit heritable traits that contribute to their replication success.

This theorem is fundamental to the population genetics approach to understanding evolution. It provides us with three necessary conditions for change due to selection:

(i) a source of variation;
(ii) a method of replication;
(iii) covariance between variants and their replication success.

Intuitively, cultural characteristics seem to fit these criteria. However, our intuitive impressions are an insufficient basis for a rigorous model of cultural evolution. We argue that the detection and description of *units* of selection is an essential first step towards applying models of natural selection to the realm of culturally transmitted information. All three conditions listed above revolve around a single more fundamental assumption: the presence of differentially replicating units. It is to these units of selection that we shall now bring our attention.

2. Units of Selection

Units of selection, defined as those patterns which differentially replicate, are essential to any evolutionary model. While some transmission-based approaches to cultural change may be workable without the assumption of any sort of cultural particle (Boyd & Richerson, 1985), recent work in cultural evolution models usually assumes that some sort of unit exists (Findlay, 1992; Laland *et al*., 1995). We argue that attention to the problem of the units of selection is a task essential to the understanding of the process of cultural evolution. Much confusion in evolutionary biology has been caused by a vague conception of the units of selection (Williams, 1966), and much of the literature on the topic is more philosophical than empirical (Lloyd, 1989; Walter, 1991; Sober, 1992; Sober & Wilson, 1994; Hill, 1994). While we agree that a philosophical analysis can help direct us towards asking the right questions, units of selection should be induced empirically not deduced a priori. We must stress that as natural selection is a hierarchical theory we cannot claim to solve the problem of what is the *sole unit* of selection, but we look for *a unit* (or set of units) of selection at an appropriate level. Under different circumstances and in different systems, the units of selection may change or operate in parallel. In evolving systems, including biological ones, selection may simultaneously favor different replicators at different interacting levels of selection (Breden & Wade, 1989; Breden & Hausfater, 1990).

PREVIOUS APPROACHES

The primary approaches to modeling cultural evolution skirt the issue of replicators and go on to develop models of the process assuming that there are units with which to work. Dawkins (1976) introduces the term "meme" and suggests as examples "tunes, ideas, catch-phrases, clothes fashions, ways of making pots or of building arches". This throws many things into the definition and does not focus on any particular unit of selection. Cavalli-Sforza & Feldman (1981) describe cultural characters as "second order organisms" and define them very loosely. Boyd & Richerson (1985), while they primarily make use of particulate models, argue that particles are not a necessary part of their theory. Durham (1991: p. 420) is the most clear on the problem of units of selection: "I have therefore assumed (1) that both systems [biological and cultural] can be divided into recognizable subunits of transmission and inheritance; (2) that within all populations there are sources of variation in these units, sources that create alternative forms at least occasionally; and (3) that there exist one or more mechanisms of transmission through which these units are conveyed among the individuals of a population . . . Assumption 1 is probably the most important and most controversial of the set." Lumsden & Wilson (1981) make an attempt at defining the *culturegen*, their closest equivalent to a unit of selection. They claim that culturegens are sets of cultural traits that are polythetically similar. This leads the way for the numerical taxonomy-based approach that we champion here.

We note from the onset that the question of the detection of replicators is complicated by the particularities of the material in which they are found. We seek cultural replicators without direct access to the content of human minds. Instead, we have available the interactors (Hull, 1988) or vehicles (Dawkins, 1982) which somehow reflect mental replicators. Our approach is to use numerical taxonomic tools to reveal latent semantic structures that we take as indicative of an underlying replicating pattern. These patterns, which lie in what Williams (1992) calls the codical domain, are visible to us as the outcome of certain human behaviors. To our knowledge nothing like a "memory particle" has ever been described. However, the validity of our approach is not dependent upon a particulate model of memory. Viable units of selection may arise from a wide variety of physical storage media. The phenotypes we describe (combinations of rare words found by a principal component analysis, see Section 3) are indicative of replicators. They are not themselves replicators. It is the information, not the matter, which replicates; and it may be found translated into a multitude of forms, some of which (e.g. text) we find easier to work with than others.

PROPERTIES OF UNITS OF SELECTION

Following Williams' (1966) definition of the gene as "that which segregates and recombines with appreciable frequency" and Dawkins' (1976, 1982) descriptions of cultural replicators, we argue the following: The appropriate units of selection will be *the largest units of socially transmitted information that reliably and repeatedly withstand transmission*. This definition describes a unit that is most likely to come under selection and thus respond through the production of adaptations. While genes are perhaps more appropriately defined as an open reading frame in the DNA, or a section of DNA that creates a single protein transcript (Watson *et al*., 1987), Williams' definition of a gene still has utility. The two important characteristics of this definition are that a unit be large enough to exhibit properties that may covary with replication success and still be small enough to have robustly developing characteristics that reappear from host to host. At this point we have little information about putative units of cultural inheritance.

Unclear perspectives on the locus of selection can confuse our understanding of evolution and cause us to waste time looking for adaptations where none are likely to exist. The process of adaptation depends upon units of selection which possess variable properties that can be modified. As these units become smaller, we assume they will provide less raw substrate on which selection can act. As units become larger, they will fall prey to two problems, both of which will cause them to be less likely to generate adaptations. First, they will become less likely to reproduce with sufficient fidelity, due to the larger number of external contingencies involved in their replication process. Second, they will be subject to fewer sorting events. Sorting events are instances where one alternative versus another is differentially replicated.

Thus, larger units (presumably replicating less frequently) will be subject to selection as a weaker force (as they undergo fewer sorting events) as well as being ineffective at responding to selection when it does occur (due to their lower replicative integrity). The size of the units will represent a trade-off between increased substrate, on which selection can act, and the twin problems of reduced selection pressure (due to fewer sorting events) and reduced effective response to selection (due to contingencies). Size in this case is broadly defined and may be measured on different scales for different systems. Implicit in this discussion of the size of units of selection is the assumption that whatever the large units may be, they are composed of the smaller units. Thus, we assume some sort of hierarchical organization. For a discussion and review of hierarchical organization schemes of cultural replicators and their parallels in biological systems see Sereno (1991).

CRITICISMS OF CULTURAL REPLICATORS

The claim that ideas are not particulate may be raised against the cultural replicator argument. While it may be true that ideas are not always best represented as particles, there are many types of ideas that do seem to fit the replicative unit model. While we may find aspects of culture that are best described as gradients of non-particulate information, the existence of easily repeated and remembered cultural elements, such as choruses, tunes, recipes, expressions, figures of speech and religious rites suggests that at least some elements of culture can be described as discrete cultural particles with tractable phylogenetic histories. At this point, the field of cultural evolution is in such a primitive state of development that comparatively simple cultural patterns such as bird song choruses (Payne *et al*., 1988; Shackell *et al*., 1988; Lynch *et al*., 1989; Gibbs, 1990; Laland, 1992) are still poorly understood. It is not a refutation of the theory that larger bodies of culture such as economic and religious systems may now reside outside our purview. Simple repeated patterns are the units of analysis for this preliminary foray into the empirical basis of cultural replicator theory.

In a critical review of cultural selectionism, Hallpike (1986, p. 46) suggests, "theories of basic units of culture do not rest on any evidence, or on any sociological theory at all, but are simply proposed because if one is trying to explain culture on the basis of a neo-Darwinian theory of natural selection, it is highly inconvenient not to have a 'unit' like the meme or culturgen”. While his proposition that there is no evidence for “units” of culture is unsubstantiated, his claim that the lack of units is “inconvenient” is understated. We argue that some unit of cultural evolution is essential for further progress in the field, and, until the units of selection in cultural evolution are formally modeled and empirically detected, the entire body of theory lies in a precarious situation.

We now propose a text analysis method based on Latent Semantic Indexing as a means of detecting those units of selection within our model cultural system: discourse on the Internet. Our approach reveals those elements of text which demonstrate replication. Later on we give evidence that these replicators differ in fitness and are therefore units of selection.

3. Text Analysis and a Cultural Unit of Selection

We have developed an analytic technique to detect units of selection within a corpus of texts. These units or cultural replicators will be sets of words which repeatedly co-occur. We argue that the repeated co-occurrence of words across texts indicates replication of the concepts signified by those word combinations. These replicating word combinations are taken to be units of selection as they fulfill the criteria of repeatable, reliable replication. Moreover, we have shown that in one example, the degree to which these replicators are expressed strongly covaries with the “reproductive success” of texts within the corpus. This offers evidence that our replicators are suitable units of selection. Not only do they have sufficient copying fidelity, but they demonstrate trait/fitness covariance. For a more complete description of this system see Best (1996, 1997).

NETNEWS

Our corpus is composed of texts posted to the USENET News (NetNews) system. NetNews is an electronic discussion system which was developed for and is supported on the Internet. NetNews originated in 1979 as a software mechanism to distribute among networked computers “bulletins, information, and data . . . items of interest such as software bug fixes, new product reviews, technical tips, and programming pointers, as well as rapid fire discussions of matters of concern to the working computer professional” (Kantor, 1986). But today it has grown into much more then a place to discuss technical topics. Discussion groups have formed along subjects ranging from science to politics to literature to various hobbies. The collection of messages over NetNews are organized into particular subject groups, called *newsgroups*. The newsgroups themselves are organized in a tree-like hierarchy which has at the root general top-level categories and moves to more specific topics as you progress towards the leaves. A newsgroup name is defined as the entire path from the top-level category through any subsequent refining categories down to the name of the group itself. Category and group names are delimited by the period symbol. Thus, “sci.biology” is the name of a scientific-oriented newsgroup devoted to general biology subjects. And “sci.biology.evolution” is a more specific group devoted to the study of evolution.

Posts to NetNews are composed of a number of fields, only a few of which are relevant here. The user creating the post is responsible for the post “body” (that is, the actual text of the message) as well as a subject line. The subject line is composed of a few words which describe what the post is about. NetNews software will attach to posted messages a number of additional fields including a timestamp and the user name of the person who created the post. A fictitious example of a post sent to the sci.bio.evolution newsgroup along with some of its header information follows:

```
From: someone@foo.net
Newsgroups: sci.bio.evolution,
alt.architecture
Subject: Meme adaptations?
Date: 26 Sep 1997 02:17:05 -0700

After reading _Dar-
win’s_Dangerous_Idea_ (Dennett) and
the ‘‘Spandrels of St. Marx’’ paper
(Gould and Lewontin) I have become
confused about the term ‘‘adap-
tation.’’ Can anyone corroborate
Dennett’s claim that church span-
drels were actually an adaptation
in that they were designed/selected
to *meme* the churchgoers with
great big church icons?
```

Note the cross-posting to another relevant newsgroup outside of the sci.* hierarchy (alt.architecture).

Posts can be either an independent message or a follow-up to a previous message. A follow-up, or “in-reply-to” message, will have special threading information in its header linking it to the previous posts to which it is a reply. This header information allows news readers to reconstruct the discussion thread.

NetNews today has grown considerably from its beginnings in the late 1970s and 1980s. With over 80000 posts arriving each day, it provides an excellent dataset for the study of cultural microevolution.

TEXT ANALYSIS MECHANISMS

Our goal is to analyse a collection of posts to NetNews in order to distill from each post those salient units of selection—that is to say, those sets of words which are replicating reliably and repeatedly and thus may be targets of selection. To reach this goal we employ a number of text retrieval techniques which read and convert each post into a vector representation based on word occurrences. We then perform a principal component analysis of these vector representations, based on Latent Semantic Indexing (Furnas, *et al*., 1988), which distill from the corpus those most statistically relevant word co-occurrences; these will be our replicating units. Finally, we propose a method to measure the *trait/replicative success* covariance of our replicators versus the posts. By demonstrating a strong covariance, we argue that some of our replicators are indeed subject to selection. These steps, then, allow us to conclude that our methods are a useful analytic technique for distilling cultural units of selection.

Given the full-text of a particular post, we wish to determine a vector-space representation (Salton & Buckley, 1988). Not all words in the text will be considered during our analysis. For example, extremely frequent words are discarded (e.g. "the", "and", "or"), and suffixes are stripped off (e.g. "computers" becomes "computer"). The resultant list of words across the entire corpus is called the *term list*. We score each document according to the frequency of occurrence of each term within its text and assign to each document/term pairing this score or *term weight*. The term weighting we use for each post is a function of the *term frequency* (simply the number of times the term occurs in the post) and the *inverse document frequency* (*IDF*) (Croft & Harper, 1979). Consider a corpus of m posts and a particular term, j, within a list of n terms. Then the IDF is given by,

$$IDF_j = \log\left(\frac{m - m_j}{m_j} \right),$$

where m_j is the number of posts across the entire corpus in which term j appears. The term weight for a document, i, and term, j, is then defined by,

$$TermWeight_{ij} = w_{ij} = \log(TermFrequency_{ij}) \cdot IDF_j.$$

Each term weight, then, is a combination of the inter- and intra-document term frequencies.

Each post, i, is now represented by a particular term vector,

$$\mathbf{r}_i = (w_{i1}, w_{i2}, \ldots, w_{in}).$$

And the entire collection of m term vectors, one for each post, define the *term/document matrix*, **A**,

$$\mathbf{A} = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \cdots \\ \mathbf{r}_m \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{bmatrix}.$$

This set of steps, culminating in the term/document matrix, forms the basis for much of modern text retrieval or filtering [see for instance Frakes & Baeza-Yates (1992)]. Given these initial steps, we are now ready to compute the replicating word co-occurrences. We wish to apply a statistical technique which meets our theoretical goals of extracting elements of the text which represent replicating units of reasonable size and copy fidelity. The statistical technique we have made use of is a principal component analysis called singular value decomposition or SVD. Applying SVD to the term/document matrix finds those term co-occurrences which segregate and recombine with appreciable frequency but still with variation. Further, we claim that SVD distills those term co-occurrences which have sufficiently salient underlying semantic structures so as to be responsive to selection.

The use of SVD for text-retrieval applications was originally proposed and has been extensively studied by Susan Dumais of Bell Communications Research and her colleagues (Furnas *et al*., 1988; Deerwester *et al*., 1990; Dumais, 1992, 1993) They refer to this technique as latent semantic indexing (LSI). Peter Foltz has investigated the use of LSI in clustering NetNews articles for information filtering (Foltz, 1990). Michael Berry and co-authors have researched a variety of numerical approaches to efficiently perform SVD on large sparse matrices such as those found in text retrieval (Berry, 1992; Berry *et al*., 1993; Berry & Fierro, 1995). Our approach follows the LSI approach closely, though we use their methods in novel ways.

Overview of SVD

The SVD technique decomposes the term/document matrix into a left and right orthonormal matrix of eigenvectors and a diagonal matrix of eigenvalues. The decomposition is formalized as,

$$\mathbf{A} \approx \mathbf{A}_k = \mathbf{U}\Sigma\mathbf{V}^T = \sum_{i=1}^{k} \mathbf{u}_i \cdot \mathbf{v}_i^T.$$

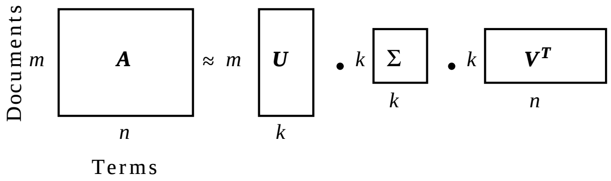

FIG. 1. Decomposition of term/document matrix into rank-k approximation.

Thus, the term/document matrix, **A**, is approximated by a rank-k decomposition, A_k; in fact the SVD technique is known to produce the *best* rank-k approximation to a low-rank matrix (Berry, 1992). Figure 1 shows graphically this decomposition.

We are only interested in the right orthonormal matrix of eigenvectors, $\mathbf{V}^T$. Each row of this matrix defines a set of terms whose co-occurrences have some statistically salient structure to them. That is, each eigenvector describes a subspace of the terms which are frequently found together. These *term-subspaces* describe a set of semantically significant associative patterns in the words of the underlying corpus of documents; we can think of each subspace as a *conceptual index* into the corpus (Furnas *et al*., 1988). For instance, in Fig. 2 we see a term-subspace which marks three terms as having significant co-occurrences, and therefore replicating together with success: "harbor", "japan", and "pearl". (Note that this term-subspace was the result of analysing a collection of military posts.) It is these term-subspaces which make up our replicators and are our putative units of selection.

Our final step is to project the original term/document matrix onto the term-subspaces by multiplying it with this right orthonormal matrix of eigenvectors. This, in effect, produces a *term-subspace/document* matrix. Each post now is represented by a collection of weights where each weight describes the degree to which a term-subspace is expressed within its post's text. Thus, each post is represented by the degree to which it expresses the morphology described by each of the term-subspace replicators. Since each of these term-subspace weights are real-valued, they define a *metric trait* for their post. We can think of the term-subspace vector representation for each post as a *memotype*.

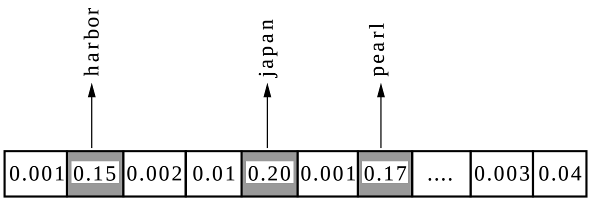

FIG. 2. Most significant weights in the vector (shaded) represent the salient terms.

TRAIT/REPLICATIVE SUCCESS COVARIANCE

We have argued that an outcome of the SVD, due to its statistical properties, is that the term-subspace represents a distinct replicator within the corpus. We described a means, through projection, of representing each post as a vector of metric traits where each trait is the degree to which one of these replicators is expressed within the post. We now demonstrate that the reproductive success of a *post* (rather than just the replicative success of each term-subspace) is a function of the degree to which certain term-subspace traits are expressed within the post. That is to say, the fact that replicators exist within the corpus does not necessarily mean that the success of a post is at all related to the degree to which a replicator is expressed. We will show that the expression of a trait is related to the success of a post by demonstrating a strong covariance between some trait's metric value and the replicative success of a collection of posts.

Threading within NetNews is designed to chain semantically similar posts together—each subsequent post is designed to be a *response* to the previous post. Therefore, we can think of each post as time moves forward as the progeny of the previous post(s). For demonstration purposes we shall consider a single example of trait/fitness covariance within a thread.

We wish to assess the replicative success of an in-reply-to thread of posts. We define success as simply the density of posts over time. So as the number of posts per time unit increases for some particular thread the posts within that thread are said to be replicating with greater success. We compute this measure of replicative success by integrating the number of posts over some fixed period of time. This amounts to simply histogramming and smoothing the timestamp data.

We have computed the trait/replicative success covariance for one thread, composed of 101 posts to the sci.skeptic newsgroup, over the period of time from September 20, 1995 to September 26, 1995. These posts made up a heated discussion about the behavior of an individual, James Smith (the name has been changed), who was a prolific and controversial poster. Thus the thread is essentially a debate on the proper customs and protocols when posting to NetNews. The term-subspace vectors where computed over a larger corpus of posts, 11 758, sent over the same period of time. Our thread of 101 posts were members of this larger corpus. We distilled

27031 terms from the corpus of texts and the SVD procedure arrived at 209 replicators. Some examples of the replicators include:

(i) algorithm, fuzzy, genetic, inference, neural;
(ii) drink, milk;
(iii) energy, solar;
(iv) chlorine, depletion, ozone, stratosphere.

We then computed the covariance of a particular metric trait with the replicative success (as defined above) of our in-reply-to thread. We chose a particular trait that demonstrated a strong covariance; though many other traits show considerable correlation. The chosen term-subspace trait described the co-occurrence of the terms, "james, smith, nazi". As mentioned, this thread consisted of a debate about the posting habits of James Smith. And not surprisingly, the quality of the discussion moved quickly to name calling (particularly, calling Smith a "Nazi").

Figure 3 shows a plot against time of the *average* degree of expression of the "james, smith, nazi" trait across time along with the thread's replicative success. The significant trait/replicative success covariance is visually apparent. As the density of posts in time went up, the average degree to which the posts expressed our particular trait also increased. In other words, when posts made greater use of the replicator, "james, smith, nazi", the number of posts per unit time to that thread increased.

We computed the covariance numerically. The normalized covariance, or *correlation coefficient*, between the measure of replicative success and the trait value was 0.7048 ($p < 0.001$).

4. Discussion and Conclusions

Our definition of a unit of selection is meant to describe units of culture that are likely strongly affected by selective forces. Any smaller units will not change rapidly if larger scale amalgamations of them replicate reliably, while any unit that is larger still will not maintain its fidelity through replication and thus will not respond effectively to selection pressure. Units of selection that break apart frequently because they are composed of unstable subunits may be under selection pressure, but they fail to respond to this selection especially if there are countervailing forces acting at a lower level where response to selection occurs more rapidly. The important characteristics of a unit of selection are that it be composed of subunits that vary and that it have characteristics that relate to its replicative success. The statistical properties of the SVD mean that term-subspaces clearly meet the first criterion. Our results regarding the covariation of thread density with term-subspace expression suggest that they meet the second as well.

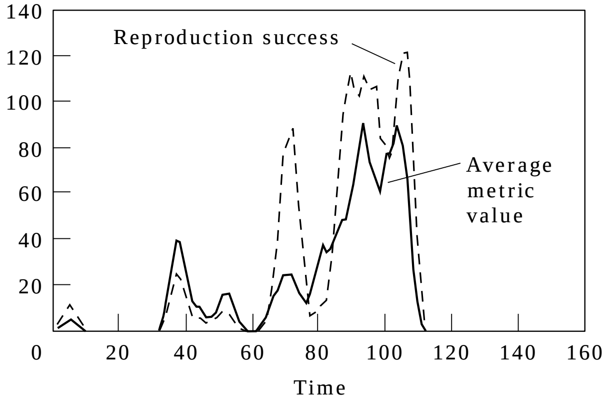

FIG. 3. The reproductive sucess of a thread of posts, computed as post density against time, covaries with the average real-value degree of expression of a replicator.

The term-subspace is composed of a number of terms that when found together in a document represent the occurrence of a particular "cognitive motif". We assay for the presence of a subspace in a manner that is similar to the methods used to detect and separate proteins (gene products) from solution. Proteins are large complex molecules that have many chemical properties, some of which are highly variable in the face of reaction conditions. This is what produces the versatility that gives them a central role in living processes. Certain elements of a protein, epitopes, are regions which have a known binding affinity. In order to detect a protein one can construct an affinity column that contains a ligand that binds with a particular epitope. When the solution is passed through the column, those structures containing the epitope, or a close analog, are stuck to the column. Similarly, our matrix decomposition can be considered to screen various posts for the degree to which they "bind" to a particular cluster of rare co-occurring words.

In this preliminary study not only do we screen the text, but we also go through the exercise of deciding on appropriate epitopes to screen for. Thus we use a recognizable epitope-like structure, shared rare words, to assay for the presence of a more ephemeral conceptual replicator. The analogy between concepts and proteins is intriguing as in both cases they are the result of a decoded message (from text or DNA) into a poorly understood structure (an idea or a protein through an RNA intermediate) that may eventually result in the replication of the message that was translated.

We argue that in the NetNews system term-subspaces are appropriate units of selection. They are the largest clusters of elements that are found repeatedly throughout the dataset. We also find that their occurrence covaries with one measurement of fitness: relative abundance.

It is important to note that these units of selection, the term-subspace word sets, are derived from the data and were not categories constructed by the investigators. The term-subspaces emerge from the data when treated with a theory-neutral statistical technique. Future analysis of the nature of the term-subspaces may provide us with a more qualitative distinction among posts. Ideally we would like to produce an affinity column model that is able to score posts for the presence or absence of predefined epitopes. This will require only minor changes in our current methods, but much more data.

In this paper we have argued that future progress in the area of cultural evolution requires more empirical study of several basic assumptions of evolutionary models. One key assumption of these is that culture contains particulate fragments which can act as units of selection. We argue that a focus on the units of selection in cultural evolution is essential as current work glosses over the issue. Our text analysis method describes sets of words that co-occur with frequency in posts to NetNews on the Internet. We claim that this technique finds markers of reliably and repeatably replicating cultural units. We show that in one case the expression of a cultural replicator covaries with a measure of the reproductive success of a post. Thus, we conclude that natural selection does and will occur in human-created environments such as the Internet. To what end we cannot as yet predict.

We wish to thank the Santa Fe Institute Complex Systems Summer School where these memes first germinated. David Christini provided early ideas and inspiration. We are grateful to the two anonymous reviewers for their helpful critiques. We also thank Jack Driscoll, Felix Breden, and Eric Minch for their close readings of this manuscript. The first author thanks NSERC and Simon Fraser University for their support of the work.

REFERENCES

Berry, M. W. (1992). Large-scale sparse singular value computations. *Int. J. Supercomputer Applications* **6,** 13–49.

Berry, M., Do, T., O'Brien, G., Krishna, V. & Varadhan, S. (1993). SVDPACKC (Version 1.0) User's Guide. University of Tennessee Computer Science Department Technical Report, CS-93-194.

Berry, M. W. & Fierro, R. D. (1995). Low-Rank Orthogonal Decompositions for Information Retrieval Applications. University of Tennessee Computer Science Department Technical Report, CS-95-284.

Best, M. L. (1996). An ecology of the net: message morphology and evolution in NetNews. Massachusettes Institute of Technology, Media Laboratory, Machine Understanding Technical Report, 96-001.

Best, M. L. (1997). An ecology of text: using text retrieval to study alife on the net. *J. Artif. Life*. (Submitted).

Boyd, R. & Richerson, P. J. (1985). *Culture and the Evolutionary Process*. Chicago: University of Chicago Press.

Breden, F. & Hausfater, G. (1990). Selection within and between social groups for infanticide. *Am. Natur*. **136,** 637–688.

Breden, F. & Wade, M. (1989). Selection within and between kin groups of the imported willow leaf beetle. *Am. Natur*. **134,** 35–50.

Cavalli-Sforza, L. & Feldman, M. (1981). *Cultural Transmission and Evolution*: *A Quantitative Approach*. Princeton: NJ: Princeton University Press.

Cavalli-Sforza, L., Feldman, M. et al. (1982). Theory and observation in cultural transmission. *Science* **218,** 19–27.

Cloak, F. (1973). Is a cultural ethology possible? *Hum. Ecol*. **3,** 161–182.

Croft, W. B. and Harper, D. J. (1979). Using probabilist models of document retrieval without relevance information. *Documentation*, **35**(4), 285–295.

Darwin, C. (1859). *On the Origin of Species*. London: Murray.

Dawkins, R. (1976). *The Selfish Gene*. New York: Oxford University Press.

Dawkins, R. (1982). *The Extended Phenotype*. San Francisco: WH Freeman.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. (1990). Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci*. **41**(6), 391–407.

Dumais, S. T. (1992). LSI meets TREC: A status report. In: *The First Text Retrieval Conference* (*TREC*-1) (Harman, D., ed.) pp. 500–207. NIST Special Publication.

Dumais, S. T. (1993). Latent semantic indexing (LSI) and TREC-2. In: *The Second Text Retrieval Conference* (*TREC*-2) (Harman, D., ed.) pp. 500–215. NIST Special Publication.

Durham, W. (1991). *Coevolution*. Stanford: Stanford University Press.

Findlay, C. S. (1992). Phenotypic evolution under gene-culture transmission in structured populations. *J. theor. Biol*. **156,** 387–400.

Fisher, R. A. (1912). Social selection. Unpublished paper read to the Cambridge University Eugenics Society.

Foltz, P. W. (1990). Using Latent Semantic Indexing for Information Filtering. In: *Proceedings of the* 5*th Conference on Office Information Systems*. ACM SIGOIS Bulletin vol. 11, issues 2,3.

Frakes, W. B. & Baeza-Yates, R. eds (1992). *Information Retrieval*: *Data Structures and Algorithms*. Englewood Cliffs, NJ: Prentice Hall.

Furnas, G. W., Deerwetser, S., Dumais, S. T., Landauer, T. K., Harshman, R. A., Streeter, L. A. and Lochbaum, K. E. (1988). Information retrieval using a singular value decomposition model of latent semantic structure. In: *Proceedings of the* 11*th International Conference on Research and Development in Information Retrieval* (*SIGIR*). New York, ACM.

Gibbs, H. (1990). Cultural evolution of male song types in Darwin's medium ground finches, Geospiza fortis. *Anim. Behav*. **39,** 253–263.

Hallpike, C. (1986). *The Principles of Social Evolution*. Oxford: Clarendon Press.

Hewlett, B. & Cavalli-Sforza, L. (1986). Cultural transmission amongst aka pygmies. *Am. Anthropol*. **88,** 922–934.

Hill, J. (1994). Units of selection in organic and sociocultural evolution. *Social Biol. Hum. Affairs* **59,** 10.

Holland, J. (1975). *Adaptation in Natural and Artifcial Systems*. Ann Arbor: University of Michigan Press.

Hull, D. L. (1988). *Science as a Process*. Chicago: University of Chicago Press.

Laland, K. (1992). A theoretical investigation of the role of social transmission in evolution. *Ethol. Sociobiol.* **13,** 87–113.

Laland, K. N., Kumm, J. & Feldman, M. W. (1995). Gene-culture coevolutionary theory: a test case. *Curr. Anthropol.* **36**(1), 131–156.

Lewontin, R. C. (1970). The units of selection. *Ann. Rev. Ecol. System.* **1,** 1–18..

Lloyd, E. (1989). A structural approach to defining units of selection. *Philos. Sci.* **56,** 395.

Lumsden, C. & Wilson, E. (1981). *Genes, Mind and Culture.* Cambridge: Harvard University Press.

Lynch, A., Plunkett, G. M., Baker, A. J. & Jenkins, P. F. (1989). A model of cultural evolution of chaffinch song derived with the meme concept. *Am. Natur.* **133**(5), 634–653.

Payne, R. B., Payne, L. L. & Doehleri, S. M. (1988). Biological and cultural success of song memes in indigo buntings. *Ecology* **69**(1), 104–117.

Plotkin, H. (1994). *Darwin Machines and the Nature of Knowledge.* Cambridge: Harvard University Press.

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Inf. Proc. Management.* **24**(5), 513–523.

Schuster, P. & Sigmund, K. (1983). Replicator dynamics. *J. theor. Biol.* **100,** 533–538.

Sereno, M. I. (1991). Four analogies between biological and cultural linguistic evolution. *J. theor. Biol.* **151,** 467–507.

Shackell, N., Lemon, R. et al. (1988). Song similarity between neighboring American redstarts (Setophaga ruticilla): a statistical analysis. *Auk* **105,** 609–615.

Sober, E. (1992). Screening-off and the units of selection. *Philos. Sci.* **59,** 142.

Sober, E. & Wilson, D. (1994). A critical review of philosophical work on the units of selection problem. *Philos. Sci.* **61,** 534.

Walter, D. (1991). The units of selection and the bases of selection. *Philos. Sci.* **58,** 417.

Watson, J. (1976). *Molecular Biology of the Gene.* Menlo Park: Benjamin Cummings.

Williams, G. C. (1966). *Adaptation and Natural Selection.* Princeton: Princeton University Press.

Williams, G. C. (1992). *Natural Selection.* Oxford: Oxford University Press.