Analysis of Correlations Between Sites in Models of Protein Sequences

B. G. Giraud
Alan S. Lapedes
Long Chang Liu

SFI WORKING PAPER: 1998-10-092

SANTA FE INSTITUTE

Analysis of Correlations Between Sites in Models of Protein Sequences

B.G.Giraud

Service Physique Théorique, DSM, C.E. Saclay, 91191 Gif/Yvette, France

Alan Lapedes

Theoretical Division, Los Alamos National Laboratory

The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM 87501

and

Lon Chang Liu

Theoretical Division, Los Alamos National Laboratory

Abstract A criterion based on conditional probabilities, related to the concept of algorithmic distance, is used to detect correlated mutations at noncontiguous sites on sequences. We apply this criterion to the problem of analyzing correlations between sites in protein sequences, however the analysis applies generally to networks of interacting sites with discrete states at each site. Elementary models, where explicit results can be derived easily, are introduced. The number of states per site considered ranges from two, illustrating the relation to familiar classical spin systems, to twenty states, suitable for representing amino acids. Numerical simulations show that the criterion remains valid even when the genetic history of the data samples (e.g. protein sequences), as represented by a phylogenetic tree, introduces non-independence between samples. Statistical fluctuations due to finite sampling are also investigated and do not invalidate the criterion. A subsidiary result is found : the more homogeneous a population, the more easily its average properties can drift from the properties of its ancestor.

Keywords Proteins, DNA, RNA, correlations, correlation at a distance, mutations, secondary structure, entropy, spin model, statistical dependence.

Submitted for publication to *Physical Review E*. Requests for preprints should be sent to B.G.Giraud

1. Introduction

In a previous paper [1] the covariation of mutations at noncontiguous sequence sites in the V3 loop of the HIV-1 virus was analyzed via two criteria taken from the theory of information: "mutual information" M and "information content" I. Related work may be found in [2], [3], [4], [5], [6], [7] and [8]. Extensions to the work in [1] were presented in [9] in order to address phylogenetic effects and effects of correlation at a distance. Correlation at a distance, a term familiar in analysis of spin systems [10], refers to chains of correlations between directly interacting sites, inducing correlations between sites "at a distance" that do not directly interact. The discovery of causal links between events occuring at seemingly separated sites of genetic sequences can be of great practical and theoretical importance, whether

such links indicate geometrical properties or more profound interactions. The purpose of this paper is to revisit this problem, in order to validate suggested criteria for the identification of such links via detailed simulation, and to investigate new criteria. In particular, one must be aware of statistical biases, because e.g. the mutual information M (a measure of correlation [11]) is a semi-positive definite quantity, and can only be overestimated by fluctuations. Moreover, members of a population with a common ancestry are, by definition, not statistically independent, and it is necessary to disentangle true correlations from the spurious ones which only reflect common ancestry of genetic sequences. Finally, causal links are not necessarily reciprocal, and criteria are needed which indicate whether a site i influences a site j while j is indifferent to i. Last but not least, if i influences j which in turn influences k, a correlation between i and k is likely to be detected, and it is useful to distinguish this "chain effect" from a direct causal link.

This paper is organized as follows. Section 2 gives a brief reminder of criteria, taken from the theory of information and probability calculus, to be used here. Then Section 3 describes an elementary two-state spin model for the validation of such criteria. The results of numerical simulations derived from this model are given in Section 4, with preliminary conclusions. Section 5 describes how the model can provide further conclusions about statistical fluctuations. Generalizations to a more biological model, where the spin can take, e.g., 20 values, to represent amino acids, are the subject of Section 6, and is further addressed in [12]. Finally Section 7 contains discussion and conclusions.

2. Criteria

Consider $\mathcal{M}$ sequences, labeled $a, b...$ of N sites, labeled $i, j...$ with each site carrying a "spin" (e.g. a base, an amino acid) which can take S values labeled $s, t...$ etc. (All labels may be understood here as positive integers, running from 1 up to their maximum ranges.)

We look for correlations between two given sites i and j. For that purpose we look at each sequence a in turn, and see whether the spin value at site i is s and the spin value at site j is t. After all $\mathcal{M}$ individual sequences have been seen, this defines a number N_{st}^{ij} of occurrences, and associated two-site frequencies, estimators of "two-site probabilities"

$$p_{st}^{ij} = \frac{N_{st}^{ij}}{\mathcal{M}}, \tag{1}$$

with $\sum_{st} p_{st}^{ij} = 1$. Marginal "one-site probabilities" p_s^i are found as traces upon two-site probabilities,

$$p_s^i = \sum_t p_{st}^{ij}. \tag{2}$$

This trace operation eliminates any dependence upon the second site, j.

The mutual information for pair ij may be defined by the following difference of entropies [11]

$$M^{ij} = \sum_{st} p_{st}^{ij} \, \mathrm{Log} \left(\frac{p_{st}^{ij}}{p_s^i p_t^j} \right) = S^i + S^j - S^{ij}, \tag{3}$$

where one recognizes the one-site entropies at sites i and j and the two-site entropy. This quantity needs a closer appraisal, however, first because it is symmetric under an exchange between i and j, and also because of the comparison between the following two models :

- model 1, a "ferromagnetic", for which all sequences show all spins up, at both sites i and j, namely ++ everywhere,

- model 2, another ferromagnetic, in a "racemic" mixture however, namely half of the sequences have spins up at both sites i and j and the other half of the sequences have spins down at both sites, all told half ++ and half $--$.

The numbers which describe the situation are

- model 1 : $p^i_+ = 1$, $p^i_- = 0$, hence a one-site entropy $S^i = 0$ at site i, then $p^j_+ = 1$, $p^j_- = 0$, hence again a one-site entropy $S^j = 0$ at site j, and finally $p^{ij}_{++} = 1$, $p^{ij}_{+-} = p^{ij}_{-+} = p^{ij}_{--} = 0$, hence a two-site entropy $S^{ij} = 0$. All told, a mutual information indicator $M = 0$;

- model 2 : $p^i_+ = p^i_- = p^j_+ = p^j_- = 1/2$, hence $S^i = S^j = \mathrm{Log}2$, then $p^{ij}_{++} = p^{ij}_{--} = 1/2$, $p^{ij}_{+-} = p^{ij}_{-+} = 0$, hence $S^{ij} = \mathrm{Log}2$. All told, $M = \mathrm{Log}2$.

Both models share the important property that sites i and j have strictly the same spin, which is a type of link we are interested in. The two differing values found for M unfortunately hides the similarity of the models. This similarity can be quantified by an alternate measure, related to algorithmic information theory [13] introduced below.

By definition, if i influences j, the conditional probability of finding t at j when i carries s is

$$p(jt|is) = \frac{p^{ij}_{st}}{p^i_s}. \tag{4}$$

Hence the following "conditional entropy" S^{ij}_s describes the influence upon j when i carries s,

$$S^{ij}_s = -\sum_t p(jt|is)\mathrm{Log}[p(jt|is)]. \tag{5}$$

Since the situation at i most often is not "pure", because more than one value of s is found from one sequence to another, the indicator to be used as a criterion is, logically, the following sum of weighted conditional entropies,

$$\Delta(j|i) = \sum_s p^i_s S^{ij}_s = S^{ij} - S^i. \tag{6}$$

This indicator, when symmetrized as $\Delta(j|i)+\Delta(i|j)$, gives the algorithmic information Δ^{ij} already described elsewhere [13]. It is trivial to verify that, for the two abovementionned "ferromagnetic" models, the two values of the conditional entropy agree, $\Delta(j|i) = 0$. As a weighted sum of conditional entropies, this indicator has a satisfactory intuitive interpretation. The value 0 found in the special case of these two models does correspond to a perfect "link" between i and j. It must be noticed, however, that $\Delta(j|i)$ is a semi-positive definite quantity like M^{ij}, hence fluctuations due to finite sample effects in the estimation of the probabilities involved in the definition of this indicator will result in likely overestimations. It will also be noticed that the lack of symmetry of $\Delta(j|i)$ under the exchange of i and j is useful if j does not influence i.

A slightly different approach was used in [1], where one defined

$$I_s^{ij} = p_s^i (S^j - \mathcal{S}_s^{ij}). \tag{7}$$

The weighting by p_s^i is the same, but $\mathcal{S}_s^{ij}$ comes with an opposite sign, and a contribution by S^j ensures the sum rule $\sum_s I_s^{ij} = M^{ij}$. In this paper, the indicator $\Delta(j|i)$, whose interpretation seems to be easier, is studied, together with M^{ij}. The next Section, Sec.3, describes a test of this indicator via a schematized model of genetic evolution.

3. Elementary model

To validate the proposed observables $\Delta(j|i)$ the following simple model is introduced. It consists of seven rules,

Maximum Simplicity Rule - There are only spins up and spins down. Spin up is coded as $+1$, or, when matrix indices are needed, by index 1. Spin down is coded as -1 or index 2.

Link Isolation Rule - Again for the sake of simplicity, no chain effect is allowed. If i influences j, then j cannot influence any other site, not even i.

Thermal Flip Rule - Still for the sake of simplicity, probabilities of mutations do not depend on time. For each time interval, for each sequence, a random number generator (RNG) generates N integers between 1 and N, with a flat distribution, hence each site in each individual sequence is statistically triggered once. Once triggered, this site may flip its spin, with a fixed probability α (thermal flip).

Influenced Flip Rule - However, if i influences j, the triggering of j induces the calculation of a modified probability $(1-\gamma st)\alpha$ for the flip, where γ is time independent. If γ is, *e.g.*, a positive number, the flip probability is thus smaller when the spins s and t at i and j, respectively, are parallel. (Alternately, rather than using a parameter γ, one may as well set two distinct, arbitrary probabilities for spin flips creating and destroying ferromagnetism.)

Binary Tree Rule - An N site "ancestor", "generation #1", is selected at random. Its sites evolve under the Thermal or Influenced Flip Rules for T time intervals (during which the RNG thus generates TN integers between 1 and N). Then it duplicates into 2 identical copies, making generation #2. The sites of these descendents evolve for again T time intervals (thus $2TN$ site labels are generated by the RNG), at the end of which each member of this generation duplicates into identical twins. The process is stopped at the end of generation #G, with a population $\mathcal{M} = 2^{G-1}$.

Star Rule - For comparison with the statistical properties of the "tree population", where successive duplications into identical twins obviously create "genetically correlated" degrees of freedom, 2^{G-1} identical copies of the same ancestor are considered initially and evolve independently for TG time intervals.

Averaging Rule - For each ancestor one calculates the values of S^i, S^{ij}, $\Delta(j|i)$, etc. as properties of the corresponding final generation, generation #G. As will be discussed later in this paper, statistical fluctuations are not negligible. It is thus necessary to average such quantities over a sampling of independent ancestors. In the next Section, Sec.4, this averaging is performed over hundreds of ancestors.

4. Numerical results from the elementary model

A priori, because of the Link Isolation Rule, it could be sufficient to consider only two sites, one independent and one influenced. Larger values of N, however, with a small proportion of interacting pairs of sites, are mandatory to provide at least an intuitive estimate of statistical fluctuations. Out of many numerical runs, the following results correspond to $N = 10$, with site 2 and 3 programmed to be influenced by site 8 and 9, respectively. It will be noticed that with $N = 10$ the number of independent ancestors is 1024, hence a few hundred ancestors ($\sim 300 - 500$) are enough for the implementation of the Averaging Rule while leaving room for some fluctuations.

Typical results are described by the following "tree" and "star" matrices $\Delta(j|i)$, displayed, respectively, at the left- and right-hand sides. (Notice that i is a row index and j a column index.) The parameters are $G = 6$, hence $\mathcal{M} = 32$, then $\alpha = 0.0015$ and $T = 20$, which give an average total probability $\sim \alpha T = 0.03$ for thermally mutating each site during each generation. For influenced spins, when triggered, we set an elementary probability 11 times larger than α to flip into a ferromagnetic situation and a strictly null probability for antiferromagnetic flips. Such a choice corresponds to a strong interaction.

$$\begin{bmatrix}
0 & 212 & 178 & 211 & 196 & 193 & 210 & 203 & 202 & 199 \\
214 & 0 & 179 & 212 & 197 & 193 & 211 & \mathbf{147} & 204 & 203 \\
214 & 213 & 0 & 212 & 200 & 195 & 213 & 205 & \mathbf{152} & 203 \\
214 & 212 & 178 & 0 & 196 & 193 & 211 & 205 & 204 & 203 \\
212 & 210 & 180 & 210 & 0 & 192 & 208 & 202 & 205 & 201 \\
213 & 211 & 179 & 211 & 197 & 0 & 210 & 204 & 203 & 205 \\
212 & 211 & 179 & 211 & 194 & 191 & 0 & 204 & 202 & 200 \\
213 & \mathbf{154} & 179 & 212 & 196 & 193 & 212 & 0 & 203 & 201 \\
211 & 211 & \mathbf{125} & 211 & 198 & 192 & 209 & 203 & 0 & 201 \\
210 & 211 & 178 & 211 & 196 & 195 & 209 & 203 & 202 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 280 & 287 & 279 & 278 & 290 & 290 & 276 & 283 & 279 \\
291 & 0 & 288 & 279 & 275 & 289 & 292 & \mathbf{179} & 282 & 279 \\
291 & 281 & 0 & 279 & 277 & 290 & 291 & 276 & \mathbf{190} & 278 \\
291 & 279 & 287 & 0 & 278 & 290 & 293 & 275 & 283 & 280 \\
292 & 278 & 288 & 281 & 0 & 289 & 291 & 276 & 283 & 279 \\
290 & 278 & 287 & 279 & 275 & 0 & 292 & 276 & 283 & 280 \\
288 & 279 & 285 & 281 & 275 & 290 & 0 & 275 & 282 & 276 \\
291 & \mathbf{182} & 287 & 278 & 277 & 290 & 291 & 0 & 282 & 279 \\
291 & 278 & \mathbf{194} & 279 & 277 & 291 & 292 & 276 & 0 & 280 \\
291 & 278 & 285 & 280 & 276 & 290 & 289 & 275 & 283 & 0
\end{bmatrix}$$

Here, both matrices are normalized by a denominator Log4, the maximum entropy of a system of two spins $\frac{1}{2}$, and, for clarity, only the integer part of $1000\Delta(j|i)/\text{Log4}$ is shown. Several results appear :

– Diagonal terms $S^{ii} - S^i$ trivially vanish. Indeed, according to the definition of N^{ij}_{st}, there is no difference at site i between counting two-site situations ss and one-site situations s, and $N^{ii}_{st} = 0$ if $s \neq t$;

– For non diagonal terms, except for a few matrix elements, $\Delta(j|i) \lesssim S^j$, as seen from the corresponding lists of one-site entropies, $\{226, 223, 189, 223, 210, 205, 223, 216, 216, 215\}$ and $\{303, 292, 299, 292, 289, 303, 305, 289, 295, 292\}$ for the tree and star cases, respectively. A slight, systematic underestimation reflects the usual overestimation of M^{ij};

– The "genetic " correlations present in the tree data give systematically smaller entropies than those for the star;

– Conversely, fluctuations for the tree case seem to be often a little stronger;

– The main result is that, for both cases, despite all sources of errors, matrix elements $\Delta(2|8)$, $\Delta(3|9)$, $\Delta(8|2)$ and $\Delta(9|3)$ stand out as smallest in their column, ensuring the detection of links;

– There is no clear hierarchy, however, between $\Delta(2|8)$ and $\Delta(8|2)$, nor between $\Delta(3|9)$ and $\Delta(9|3)$, to detect that 8 and 9 influence 2 and 3, respectively, and not the reverse;

– While sites 2 and 3 have identical properties, $\Delta(2|8)$ and $\Delta(3|9)$ differ by an amount which indicates that fluctuations are indeed not negligible;

– Sites 1 and $4-10$, which are neither influenced nor influencing, show similar columns in the tree matrix (except for the presence of $\Delta(8|2)$ and $\Delta(9|3)$, naturally). The same remark holds for the star matrix. Slight differences between such columns indicate a "normal" amount of fluctuations.

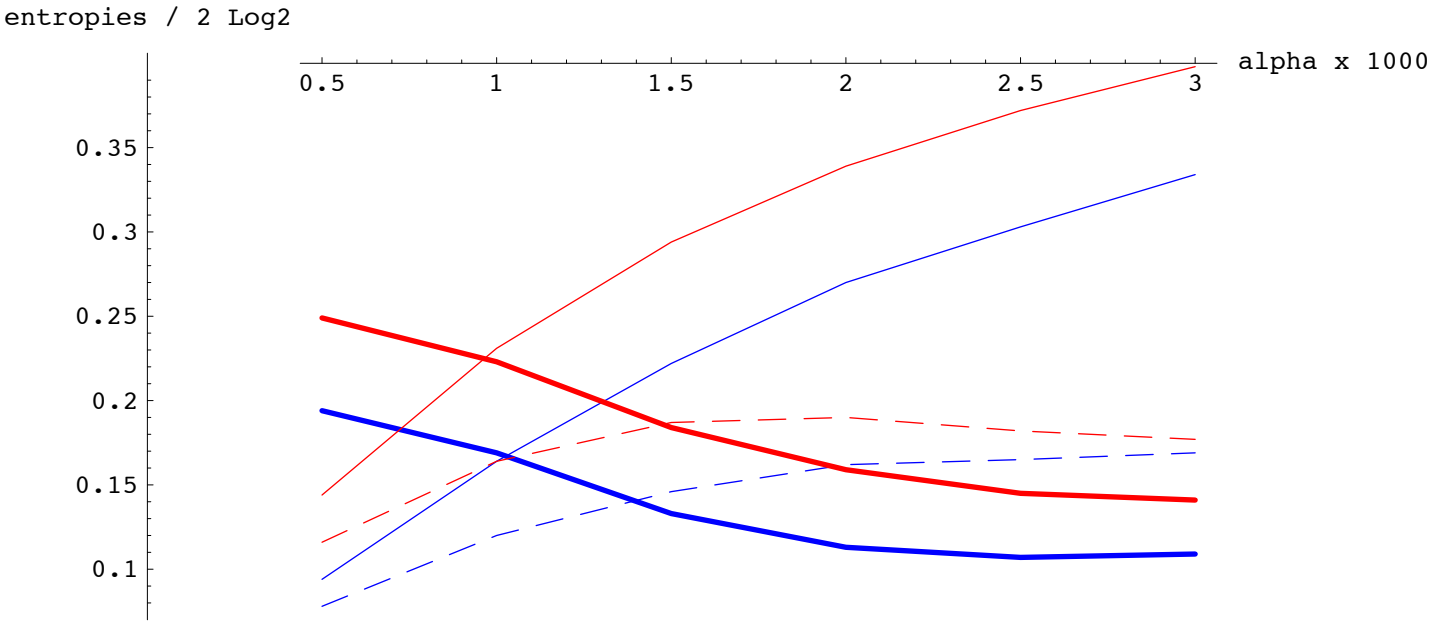
entropies / 2 Log2
alpha x 1000
0.5
1
1.5
2
2.5
3
0.35
0.3
0.25
0.2
0.15
0.1

FIG. 1. Nonreciprocal influences. Tree and star evolution properties as functions of the mutation rate $\alpha(\times 10^3)$. Light full lines: average single site entropy for isolated sites. Heavy full lines: average weighted conditional entropy for sites under influence. Dashed lines: average weighted conditional entropy for influencing sites. For each type of line, the lower, resp. upper line corresponds to tree, resp. star results.

All these conclusions are stable when one studies the influence of the rate α. This is illustrated by Fig.1, and an additional observation may be made : if i influences j, there seems to be, for both the tree and the star models, a "small α" regime, where $\Delta(j|i)$ is larger than its spurious partner $\Delta(i|j)$ in column i, and a "large α" regime, where $\Delta(j|i) < \Delta(i|j)$. The case $\alpha = .0015$ shown above was indeed chosen because of its transitional situation. It must be stressed, however, that such conclusions are valid for averages only. Fluctuations in individual trees and stars were found to create many exceptions.

We now turn to the case of reciprocal influences. Namely the Link Isolation Rule is modified and now j influences i in the same way as i influences j, all other rules, the Influenced Flip Rule in particular, remaining the same. For the sake of simplicity, we again avoid any chaining of influences : interacting pairs of sites are isolated. A few illustrative results, among many runs, are shown on Fig.2, obtained with the same parameters as Fig.1 ($N = 10$, $G = 6$, $T = 20$, "ferromagnetic flip probability" 11α, "antiferromagnetic flips" forbidden), six isolated sites, two symmetrically interacting pairs, $\{27\}$ and $\{49\}$, and 500 random ancestors. The signature for influence is again a strong minimum of $\Delta(j|i)$ as a function of i in column j.

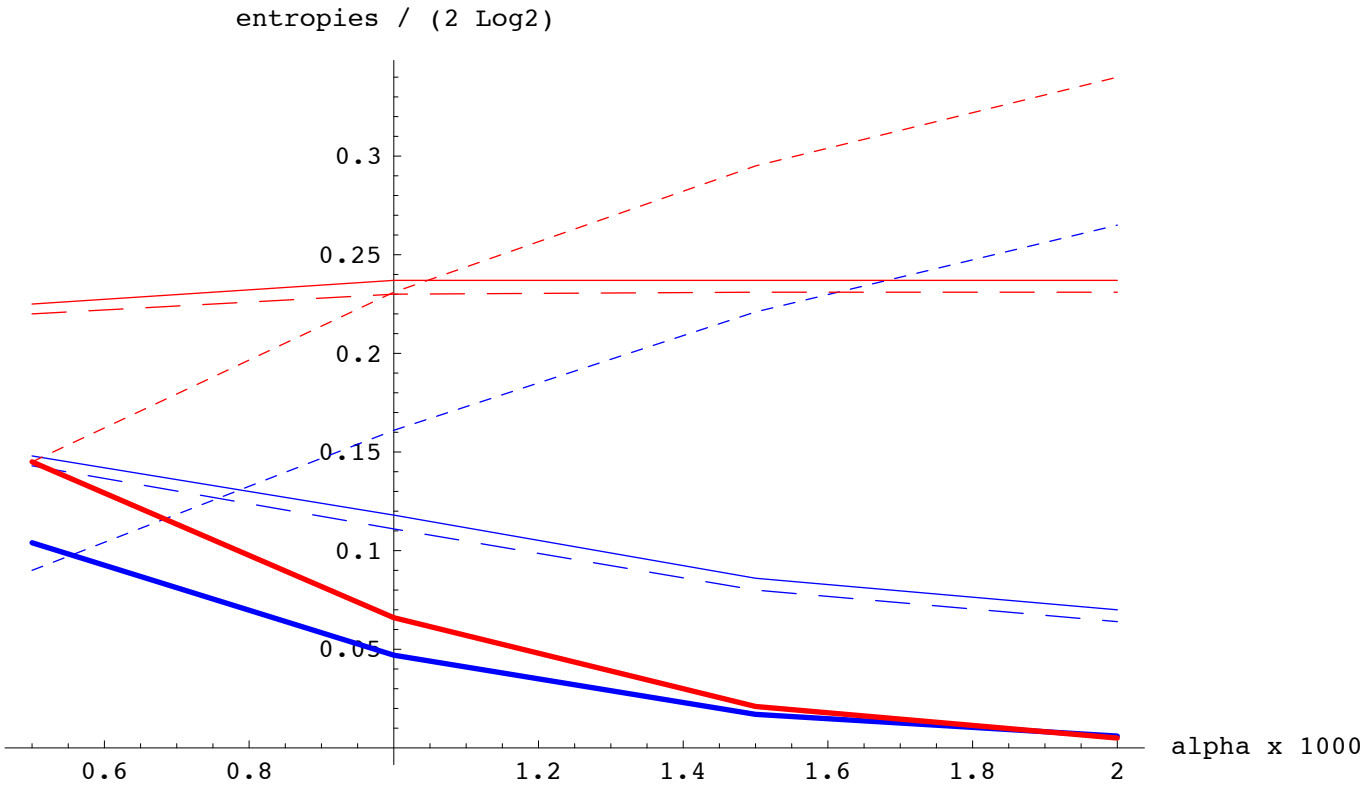
entropies / (2 Log2)
0.3
0.25
0.2
0.15
0.1
0.05
0.6
0.8
1.2
1.4
1.6
1.8
2
alpha x 1000

FIG. 2. Reciprocal influences. Tree and star evolution properties as functions of the mutation rate $\alpha(\times 10^3)$. Dotted lines: average single site entropy for isolated sites. Light full lines: average single site entropy for sites under reciprocal influence. Dashed lines: average weighted conditional entropy in those columns of matrix Δ where influence is detected. Heavy full lines: average weighted conditional entropy for sites under reciprocal influence. For each type of line, the lower, resp. upper line corresponds to tree, resp. star results.

For both the tree and the star cases, the average entropy for single isolated sites (dotted line) increases as a function of the thermal rate α, while the average conditional entropy for sites under influence (heavy line) decreases. The same trends were already shown by Fig.1, with somewhat different entropy values, however. We show on Fig.2 two additional kinds of results : as a light full line, the average single site entropy for sites under (reciprocal) influence, and, as a dashed line, the average of those matrix elements which, in *columns*, *contrast* with detected $\Delta(i|j)$. The latter point is best understood by the consideration of the following tree and star matrices, respectively,

$$\begin{bmatrix}
0 & 139 & 89 & 149 & 86 & 90 & 144 & 77 & 143 & 91 \\
88 & 0 & 87 & 147 & 85 & 86 & \mathbf{106} & 77 & 141 & 93 \\
90 & 138 & 0 & 148 & 86 & 88 & 144 & 78 & 142 & 92 \\
89 & 138 & 86 & 0 & 85 & 87 & 143 & 77 & \mathbf{102} & 91 \\
89 & 139 & 88 & 148 & 0 & 88 & 144 & 77 & 143 & 93 \\
90 & 137 & 88 & 148 & 86 & 0 & 144 & 78 & 142 & 93 \\
87 & \mathbf{101} & 87 & 146 & 84 & 86 & 0 & 76 & 141 & 91 \\
89 & 140 & 89 & 150 & 87 & 89 & 145 & 0 & 144 & 93 \\
88 & 138 & 87 & \mathbf{108} & 86 & 87 & 143 & 77 & 0 & 92 \\
88 & 140 & 88 & 148 & 87 & 89 & 144 & 77 & 143 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 219 & 135 & 219 & 141 & 138 & 218 & 136 & 219 & 140 \\
142 & 0 & 135 & 221 & 140 & 137 & \mathbf{142} & 136 & 222 & 140 \\
142 & 220 & 0 & 219 & 140 & 138 & 218 & 137 & 219 & 141 \\
142 & 223 & 135 & 0 & 140 & 138 & 221 & 137 & \mathbf{146} & 140 \\
143 & 220 & 135 & 219 & 0 & 138 & 219 & 137 & 219 & 140 \\
142 & 219 & 135 & 218 & 140 & 0 & 218 & 137 & 219 & 141 \\
142 & \mathbf{144} & 135 & 221 & 140 & 138 & 0 & 137 & 222 & 140 \\
141 & 219 & 136 & 219 & 140 & 139 & 219 & 0 & 219 & 140 \\
142 & 223 & 135 & \mathbf{146} & 140 & 138 & 222 & 137 & 0 & 140 \\
142 & 220 & 135 & 218 & 139 & 139 & 218 & 136 & 218 & 0
\end{bmatrix},$$

obtained for $\alpha = .005$. The matrix elements $\Delta(2|7)$, $\Delta(4|9)$, $\Delta(7|2)$, and $\Delta(9|4)$, which reflect the influences programmed in the run, are detected as minima in columns, not rows. Accordingly, the dashed curves in Fig.2 show the averages of such 32 "larger" numbers in columns 2, 4, 7 and 9. This is, for the tree and star cases, respectively. As an additional conclusion, for all the values of α which were considered, it is found that $\Delta(j|i) \simeq \Delta(i|j)$ when i and j are reciprocally influencing each other.

All told, the expected intensity of the signal which allows a detection of an interaction corresponds to the distance between a heavy line and the associated dashed line on Fig.2. In the range of parameters displayed there, this intensity is about three times smaller for trees than for stars. For trees, it is seen to be $\sim .05\text{Log}4 \simeq .07$, which leaves hope for successful detections in realistic cases.

About fluctuations : Severe exceptions to the conclusions drawn from averages may happen in the case of *non averaged* data. As an illustration, the following tree matrix, obtained with $\alpha = .02$ and all other parameters identical

to those used for Fig.2, refers to *one* ancestor only,

$$\begin{bmatrix} 0 & 0 & 87 & 0 & \mathbf{296} & 264 & 0 & 81 & 0 & 87 \\ 494 & 0 & 100 & 0 & 499 & 272 & 0 & 100 & 0 & 100 \\ 481 & 0 & 0 & 0 & 481 & 269 & 0 & 100 & 0 & 100 \\ 494 & 0 & 100 & 0 & 499 & 272 & 0 & 100 & 0 & 100 \\ \mathbf{292} & 0 & 83 & 0 & 0 & 261 & 0 & 86 & 0 & 86 \\ 487 & 0 & 97 & 0 & 488 & 0 & 0 & 97 & 0 & 97 \\ 494 & 0 & 100 & 0 & 499 & 272 & 0 & 100 & 0 & 100 \\ 475 & 0 & 100 & 0 & 484 & 269 & 0 & 0 & 0 & 100 \\ 494 & 0 & 100 & 0 & 499 & 272 & 0 & 100 & 0 & 100 \\ 481 & 0 & 100 & 0 & 484 & 269 & 0 & 100 & 0 & 0 \end{bmatrix} .$$

Note how much this matrix conflicts with the averaged data : during the 120 time intervals of the run, only 13 mutations happened and in particular no mutation triggered influenced flips at sites 2, 4, 7 and 9. Therefore $S^2 = S^4 = S^7 = S^9 = 0$. It must be remembered here that the semi-positive nature of the mutual information induces the automatic condition $\Delta(j|i) \leq S^j$. Hence the whole corresponding columns $2, 4, 7, 9$ vanish, and no detection is possible in such columns. Also, a spurious contrast occurs in columns 1 and 5, where $\Delta(5|1)$ and $\Delta(1|5)$ stand out as much smaller, while actually this run allowed interactions inside pairs $\{27\}$ and $\{49\}$ only. Similar aberrant cases are not infrequent in simulations of a star topology as well.

To summarize this Section, Sec.4, a *contrast* between average weighted conditional entropies was found to give a detection criteria for correlations between sites. The signal may be blurred by the noise of fluctuations, however, if the interaction between sites is weak or the sampling of ancestors is not numerous enough. The next Section, Sec.5, investigates the properties of such "noise".

5. Systematic study of fluctuations

Given the above, it is clear that, for such biological sequences, there is a non negligible risk for statistical averages, taken from an actual population, to differ significantly from true probability averages. A detailed description and understanding of this risk is in order. For that purpose, we now generate a model where fluctuations can be exhibited in a transparent way. The model is very similar, except for a few details, to that explained in Section 3. Consider again a site i, occupied by a "spin" with only two allowed values, ± 1. The basic ingredient of the model is the matrix $\mathcal{P}(s, r)$ giving the probability that, within the lifetime of one "generation", site i starting with spin r finishes with spin s. Three "generations" are considered in the model, with a common "ancestor" spin r at site i. For the star, a divergence into $\mathcal{M} = 8$ descendents is allowed from the very root of the process. For the tree, a divergence into two descendents is allowed at the root, and two additional levels of duplication are allowed. At the end of the lifetime of three "generations", the probability distribution for the spins s, t, u, v, w, x, y, z of a "star population" of 8 descendents is thus,

$$\mathcal{S}(s, t, u, v, w, x, y, z) = \mathcal{P}^3(s, r)\mathcal{P}^3(t, r)\mathcal{P}^3(u, r)\mathcal{P}^3(v, r)\mathcal{P}^3(w, r)\mathcal{P}^3(x, r)\mathcal{P}^3(y, r)\mathcal{P}^3(z, r), \tag{8}$$

where, naturally, $\mathcal{P}^3$ denotes the matrix cube of $\mathcal{P}$. In turn, the probability distribution $\mathcal{T}$ for the spins of the "tree population" of 8 descendents, see Fig.3, is easily derived from the three-index probability

$$\mathcal{Q}(\ell, m, r) = \mathcal{P}(\ell, r)\mathcal{P}(m, r) \tag{9}$$

that at site i a spin r, once duplicated, turns into spins ℓ and m as its descendents. All told one finds

$$\mathcal{T}(s,t,u,v,w,x,y,z) = \sum_{\ell mnopq} \mathcal{Q}(s,t,n)\mathcal{Q}(u,v,o)\mathcal{Q}(w,x,p)\mathcal{Q}(y,z,q)\mathcal{Q}(n,o,\ell)\mathcal{Q}(p,q,m)\mathcal{Q}(\ell,m,r). \tag{10}$$

In a transparent notation, ℓ and m are here the spins of the 2 descendents at the end of the first generation. They are followed by n, o, p and q at the end of the second generation.

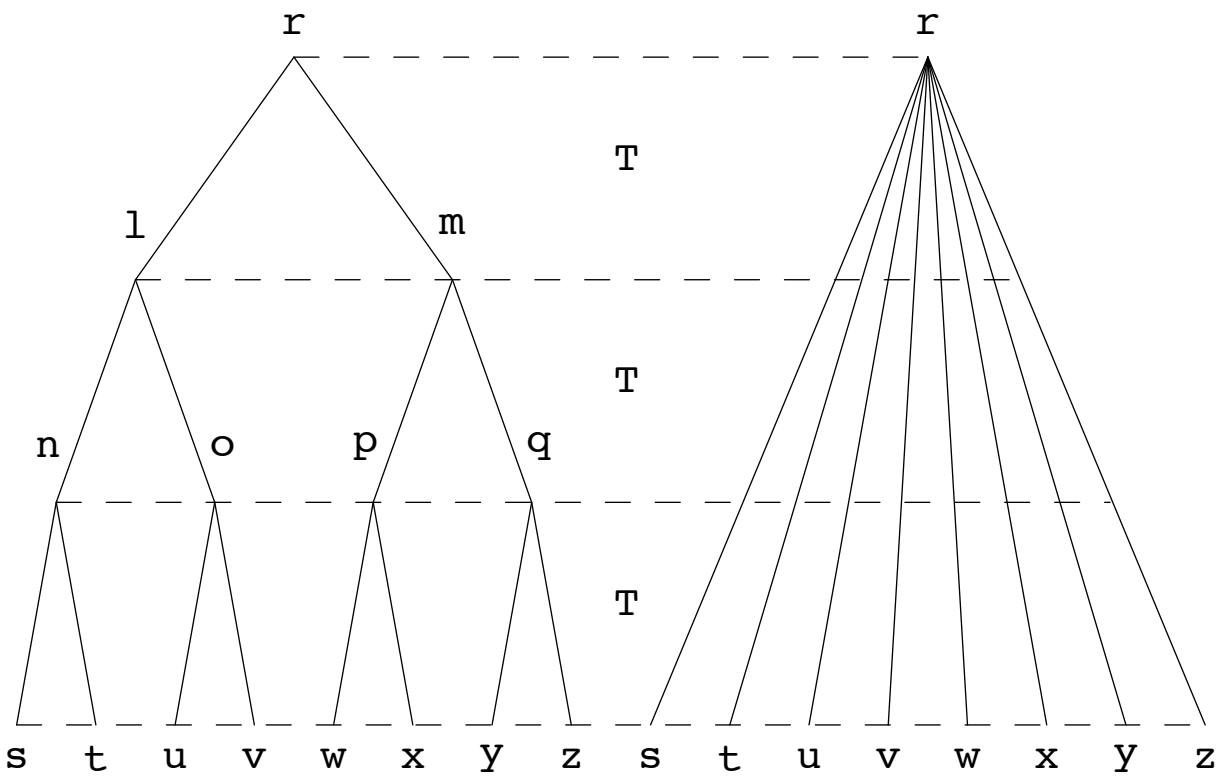

FIG. 3. Illustration of the three generation tree model used in this section, Sec.5

For the sake of simplicity, we set

$$\mathcal{P}(s,r) = \begin{bmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{bmatrix} \tag{11}$$

in the following, with a parameter ε taking on all values between 0 and 0.5. (The relation of ε with the parameter α used earlier is trivial.) Most relevant properties of the model are thus elementary functions of ε. Their study can sometimes even be reduced to polynomial operations with respect to ε. Such is indeed the case for the average value of the spin, as sampled over the population of 8 descendents and averaged over the probability distributions $\mathcal{T}$ or $\mathcal{S}$,

$$\sigma_{\mathcal{T}} = \sum_{stu\ldots z} \mathcal{T}(s,t,u,v,w,x,y,z) \frac{(s+t+u+v+w+x+y+z)}{8}, \tag{12a}$$

$$\sigma_{\mathcal{S}} = \sum_{stu\ldots z} \mathcal{S}(s,t,u,v,w,x,y,z) \frac{(s+t+u+v+w+x+y+z)}{8}, \tag{12b}$$

for the tree and star, respectively. (In a condensed notation, such integrals will be denoted as $<>_{\mathcal{T},\mathcal{S}}$ in the following.)

Here we are interested in the observable $R = (s+t+u+v+w+x+y+z)/8$, and $\sigma =< R >$, where the subscript $\mathcal{T}$ or $\mathcal{S}$ is understood. We are also interested in the fluctuations of R,

$$(\Delta R)^2 =< (R-\sigma)^2 >=< (s+t+u+v+w+x+y+z)^2 > /64 - \sigma^2, \tag{13}$$

which are given by polynomials with respect to ε. Given the same root $r = +1$ for a tree and a star, the following properties are easy to show, *i)* the average value of the spin is the same for all individuals,

$$\sigma_{\mathcal{T}} = \sigma_{\mathcal{S}} =< s >_{\mathcal{T}} =< s >_{\mathcal{S}} = \ldots =< z >_{\mathcal{T}} =< z >_{\mathcal{S}} = (1-2\varepsilon)^3, \tag{14}$$

and *ii)* for the star, correlations vanish since the branches are independent,

$$< (s-\sigma)(t-\sigma) >_{\mathcal{S}} =< (s-\sigma)(u-\sigma) >_{\mathcal{S}} = \ldots =< (y-\sigma)(z-\sigma) >_{\mathcal{S}} = 0, \tag{15}$$

while for the tree one finds $< st >=< uv >=< wx >=< yz >= (1-2\varepsilon)^2$, then $< su >=< sv >=< tu >=< tv >= < wy >=< wz >=< xy >=< xz >= (1-2\varepsilon)^4$ and finally $< sw >=< sx >= \ldots < vw >= \ldots < vy >=< vz >= (1-2\varepsilon)^6$. The relation of such overlaps with the degree of parentage of the spins is obvious. Hence the correlations $< (s-\sigma)(t-\sigma) >_{\mathcal{T}}$ at closest parentage and $< (s-\sigma)(u-\sigma) >_{\mathcal{T}}$ at next-to-closest parentage are positive definite if $0 < e < 0.5$. Any third order parentage correlation such as $< (s-\sigma)(w-\sigma) >_{\mathcal{T}}$ vanishes, as expected because there is no difference between the tree and star histories at that degree in this model. Also, obviously, any average of squared spins gives $< s^2 >= 1$ for both the tree and the star and no cross term is negative. The results for $r = -1$ are quite similar, under a replacement of $(1-2\varepsilon)$ by $(2\varepsilon - 1)$. It can be concluded that

$$(\Delta R)^2_{\mathcal{T}} = \frac{1 + (1-2\varepsilon)^2 + 2(1-2\varepsilon)^4 - 4(1-2\varepsilon)^6}{8} > (\Delta R)^2_{\mathcal{S}} = \frac{1-(1-2\varepsilon)^6}{8}, \tag{16}$$

namely the sampling of the average spin over a finite population induces a larger fluctuation for the tree than for the star. It will be noticed here that $(\Delta R)_{\mathcal{S}}$ illustrates the central limit theorem (CLT) in a transparent way. Conversely, the positive correlations brought by the tree dynamics increase the fluctuations of the average sampled spin R.

Alternate procedures are available and deserve comment, because they give different estimates of sampling fluctuations. Indeed, one may define, as a a measure of the fluctuation, the quantity

$$(\Delta\rho)^2 =< \left(s - \frac{s+t+u+v+w+x+y+z}{8} \right)^2 >, \tag{17}$$

which describes how, in each population, individual spins may deviate from the average spin. The results read,

$$(\Delta\rho)^2_{\mathcal{T}} = \frac{7-(1-2\varepsilon)^2-2(1-2\varepsilon)^4-4(1-2\varepsilon)^6}{8} < (\Delta\rho)^2_{\mathcal{S}} = \frac{7[1-(1-2\varepsilon)^6]}{8}. \tag{18}$$

Naturally one might also have considered averages of square differences between all spin pairs,

$$(\Delta\tau)^2 =< [(s-t)^2+(s-u)^2+...+(y-z)^2] > /28, \tag{19}$$

with the results,

$$(\Delta\tau)^2_{\mathcal{T}} = 2\left[1 - \frac{(1-2\varepsilon)^2+2(1-2\varepsilon)^4+4(1-2\varepsilon)^6}{7}\right] < (\Delta\tau)^2_{\mathcal{S}} = 2[1-(1-2\varepsilon)^6]. \tag{20}$$

It may be interesting to give a mechanical image of such results, Eqs.(16), (18) and (20). Consider the spins as fictitious "particles" and the "average by sampling" $R = (s+t+...+z)/8$ as their center of mass. Then the positive correlations, introduced by the tree dynamics, compress the "root mean squared relative distance" described by $\Delta\tau$ and dilate the center-of-mass fluctuation described by ΔR. This connection between the two numbers ΔR and $\Delta\tau$ may be understood as a *necessary uncertainty relation,* somewhat similar to the traditional uncertainity relation of quantum mechanics. One may make the qualitative conclusion that a reduced diversity inside a population may lead to a stronger global drift of that population. A similar intuition results from a mechanical image of $\Delta\rho$ as the dispersion of a "particle" with respect to the "center of mass". The tree dynamics tends to compress this individual dispersion, as compared to that allowed by the star dynamics independence. Accordingly, individual compressions relate to a more likely global drift.

There is no difficulty in generalizing all these considerations to models with more than $\mathcal{M} = 8$ individuals, because, obviously, "center of mass observables" imply positive signs multiplying the correlations while "relative motion observables" imply negative signs multiplying the same. The duality of such observables is systematic. In particular, while the CLT is obviously valid for $(\Delta R)_{\mathcal{S}}$, the result for $(\Delta R)_{\mathcal{T}}$, with G generations and $\mathcal{M} = 2^G$ reads,

$$(\Delta R)^2_{\mathcal{T}} = \frac{1-(1-2\varepsilon)^{2G}+\sum_{p=1}^{G-1}2^{p-1}[(1-2\varepsilon)^{2p}-(1-2\varepsilon)^{2G}]}{2^G} = 2^{-G}\left[1-\frac{\beta^G}{2}+\frac{\beta[\beta^{G-1}-1]}{2(\beta-1)}\right] = \frac{(2-\beta)(\beta^G-1)}{2^{G+1}(\beta-1)}, \tag{21}$$

with $\beta = 2(1-2\varepsilon)^2$. Since ε is small in realistic cases, an investigation of $(\Delta R)^2_{\mathcal{T}}$ in the vicinity of $\beta \lesssim 2$ is in order. For such values of β and large values of G the leading term of $(\Delta R)^2_{\mathcal{T}}$ amounts to $(2-\beta)\beta^G/2^{G+1}$, the maximum of which is reached for $\beta = 2G/(G+1)$, a value indeed hardly smaller than 2. The corresponding estimated maximum

reads $\simeq [-2^{-G} + G^G (G+1)^{-G}](G-1)^{-1}$, the asymptotic trend of which is $\simeq (eG)^{-1}$. The "tree deviation" from the CLT is thus striking, since, for comparison, $(\Delta R)^2_{\mathcal{S}}$ contains a denominator 2^G.

There is a qualitative relation between $\Delta\tau$, or $\Delta\rho$, and the one-site entropy S^i defined in Sec.2. Namely, if any one of these observables vanish, then the others vanish simultaneously. It is clear that, in turn, correlation functions between spins at site i and spins at site j would also provide an "algebraic" intuition for the behavior of the two-site entropy S^{ij}. For the sake of conciseness, however, we now investigate directly the effect of finite sampling upon the various observables S^i, S^{ij}, M^{ij} and $\Delta(j|i)$. Let $s, t, ...z$ be the 8 spins at site i and $s', t', ...z'$ be those at site j. We label the root of the three generation tree or star as r for site i, and r' for site j. No interaction is implemented between the two sites, because only "bare" fluctuations due to finite sampling are investigated here. We have thoroughly verified that, under such an independence, results are the same whether $r = r'$ or $r \neq r'$. Let $\mathcal{O}$ be any observable which is a function of all or part of the final spins $s, t, ...z'$, such as the sampled one-site entropy at site i,

$$S^i = -\left(\frac{1+R}{2}\right) \mathrm{Log}\left(\frac{1+R}{2}\right) - \left(\frac{1-R}{2}\right) \mathrm{Log}\left(\frac{1-R}{2}\right), \tag{22}$$

where $R = (s + t + ...z)/8$ is the "center of mass". The numerical results shown in the following are those averages defined by,

$$< \mathcal{O} >_{\mathcal{T}} \; = \sum_{st...z\; s't'...z'} \mathcal{T}(s, t, ...z)\, \mathcal{T}(s', t', ...z')\; \mathcal{O}, \tag{23a}$$

$$< \mathcal{O} >_{\mathcal{S}} \; = \sum_{st...z\; s't'...z'} \mathcal{S}(s, t, ...z)\, \mathcal{S}(s', t', ...z')\; \mathcal{O}, \tag{23b}$$

for the tree and star, respectively, see Eqs.(8) and (10). Of special interest are the one- and two-site, mutual and conditional entropies, as already defined in Sec.2, naturally, and the corresponding fluctuations $\Delta\mathcal{O} = \left[< \mathcal{O}^2 > - < \mathcal{O} >^2\right]^{1/2}$.

We first show, on Fig.4, the averages of the one- and two-site entropies for the three generation tree and star, respectively, in units of 2Log2, as functions of ε (multiplied by 400).

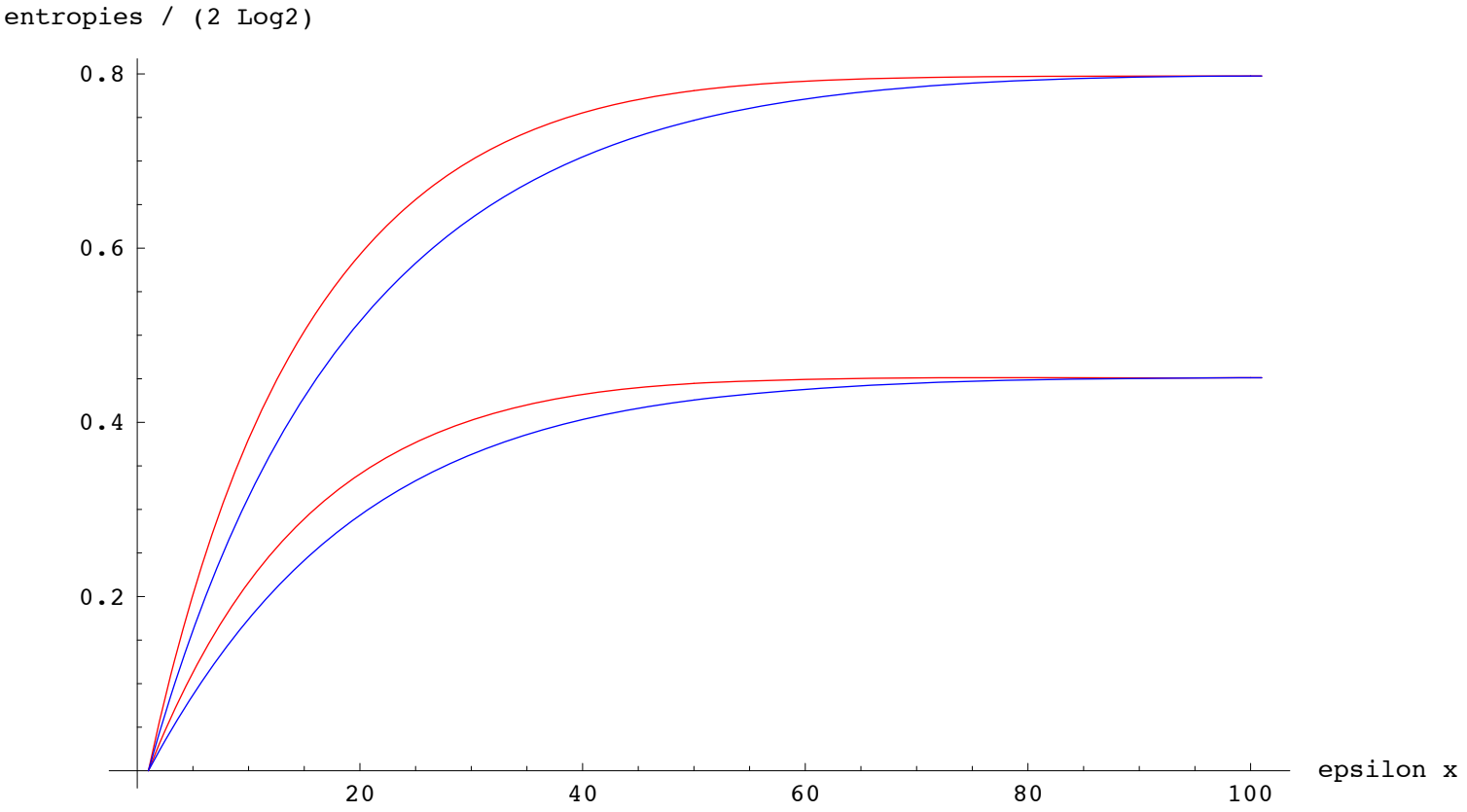
entropies / (2 Log2)
0.8
0.6
0.4
0.2
20
40
60
80
100
epsilon x

FIG. 4. Tree and star average properties after three generations. One- and two-site *sampled* entropies in units of 2Log2 as functions of the mutation rate parameter $\varepsilon(\times 400)$. Lower pair of curves: $< S^i >$. The star entropy is slightly larger than the tree one. Upper pair of curves: $< S^{ij} >$. Again $< S^{ij} >_S \gtrsim < S^{ij} >_\tau$.

A saturation is observed when $\varepsilon \gtrsim 0.2$. Both entropies are slightly larger for the star than for the tree. Since sites i and j are independent in the model, one should find $< S^{ij} >= 2 < S^i >$. However, because of the errors brought by finite sampling in a population of 8 individual sequences only, a close examination of Fig.4 shows that actually $< S^{ij} >$ is rather slightly, but systematically, smaller than $2 < S^i >$.

Then we show, on Fig.5, the averages of the mutual and the weighted conditional entropies.

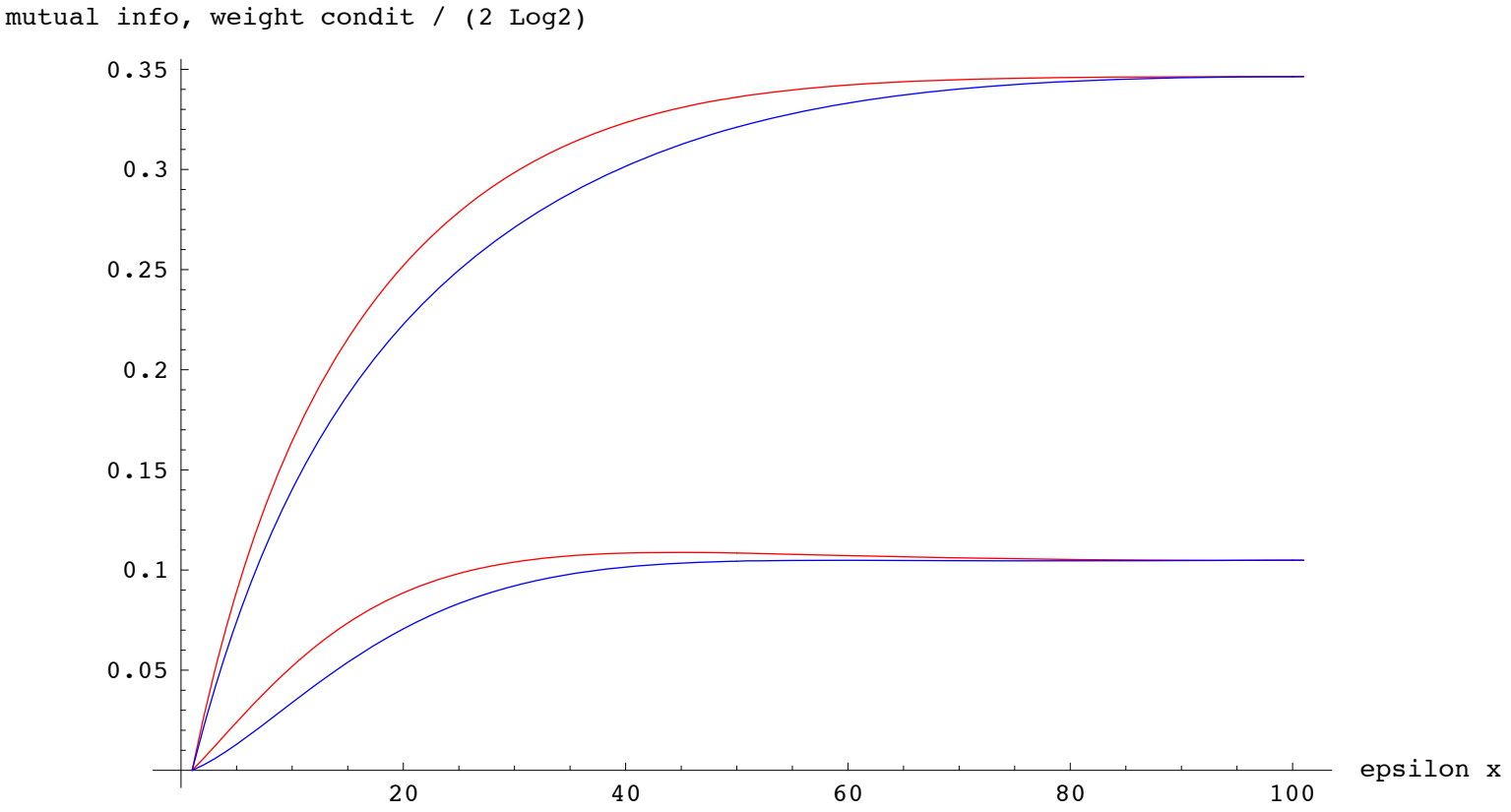

FIG. 5. Same as Fig.4, for the sampled mutual information and the sampled, weighted conditional entropy. Lower pair of curves: $< M^{ij} >$. Upper pair: $< \Delta(j|i) >$. Inside each pair, the upper curve corresponds to the star results.

Properties similar to those of Fig.4 are observed. Moreover, it must be stressed that, since the sites are independent, one might expect $< M^{ij} >= 0$ and $< \Delta(j|i) >=< S^i >$ for both the tree and the star. Clear deviations, however, due to finite sampling, are found from such predictions. Actually $< M^{ij} >$ is far from vanishing, and, furthermore, the plateau of $< \Delta(j|i) >$ on Fig.5 seriously differs from that of $< S^i >$ on Fig.4.

We now turn to the fluctuations $\Delta S^{ij}_{\mathcal{T}}$, $\Delta S^{ij}_{\mathcal{S}}$, $\Delta S^{i}_{\mathcal{T}}$, $\Delta S^{i}_{\mathcal{S}}$, shown on Fig.6, in that order, from top to bottom.

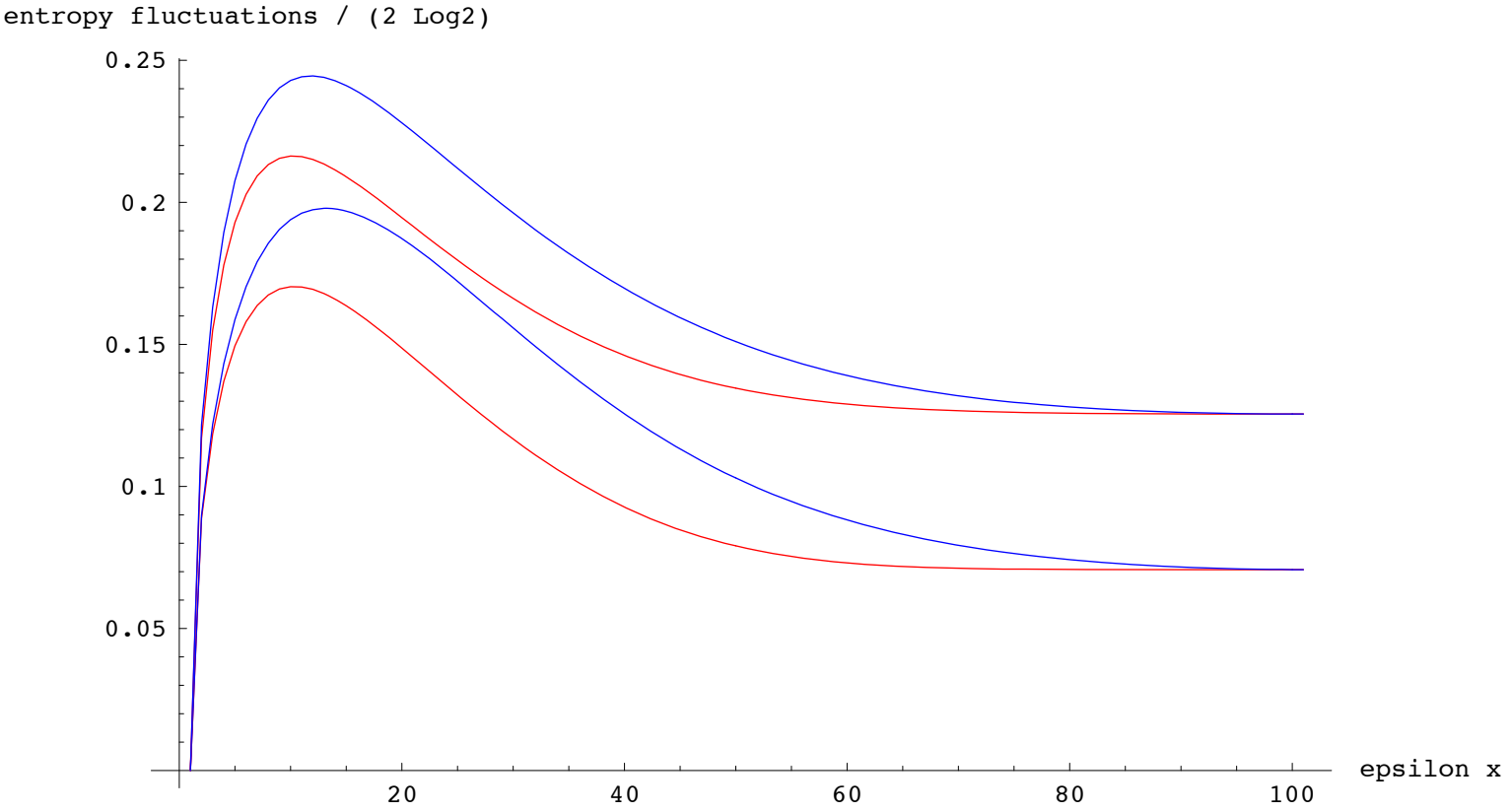
entropy fluctuations / (2 Log2)
0.25
0.2
0.15
0.1
0.05
20
40
60
80
100
epsilon x

FIG. 6. Same as Fig.4, for the fluctuations of S^i and S^{ij}. Lower pair of curves: ΔS^i. Upper pair: ΔS^{ij}. Inside each pair, the upper curve now corresponds to the tree results.

It is seen that such fluctuations are almost as large for one-site entropies as for two-site ones. Moreover, their order of magnitude can be almost as large as a one-site average entropy itself, as shown by the values reached when $\varepsilon \simeq 0.04$. This shows how the estimation of an entropy over a small population can be misleading. Two differences with Fig.4 are seen, *i)* the fluctuations are not monotonic functions of ε, and *ii)* tree fluctuations are larger than star ones, while tree entropies were smaller. Hence $\Delta S/ < S >$ is larger for trees than for stars. In terms of relative rather than absolute errors, statistical sampling from finite populations driven by evolutionay dynamics down a tree demands special caution.

Finally Fig.7 shows the behaviors of the fluctuations ΔM^{ij} (bottom pair of curves) and $\Delta[\Delta(j|i)]$ (upper pair).

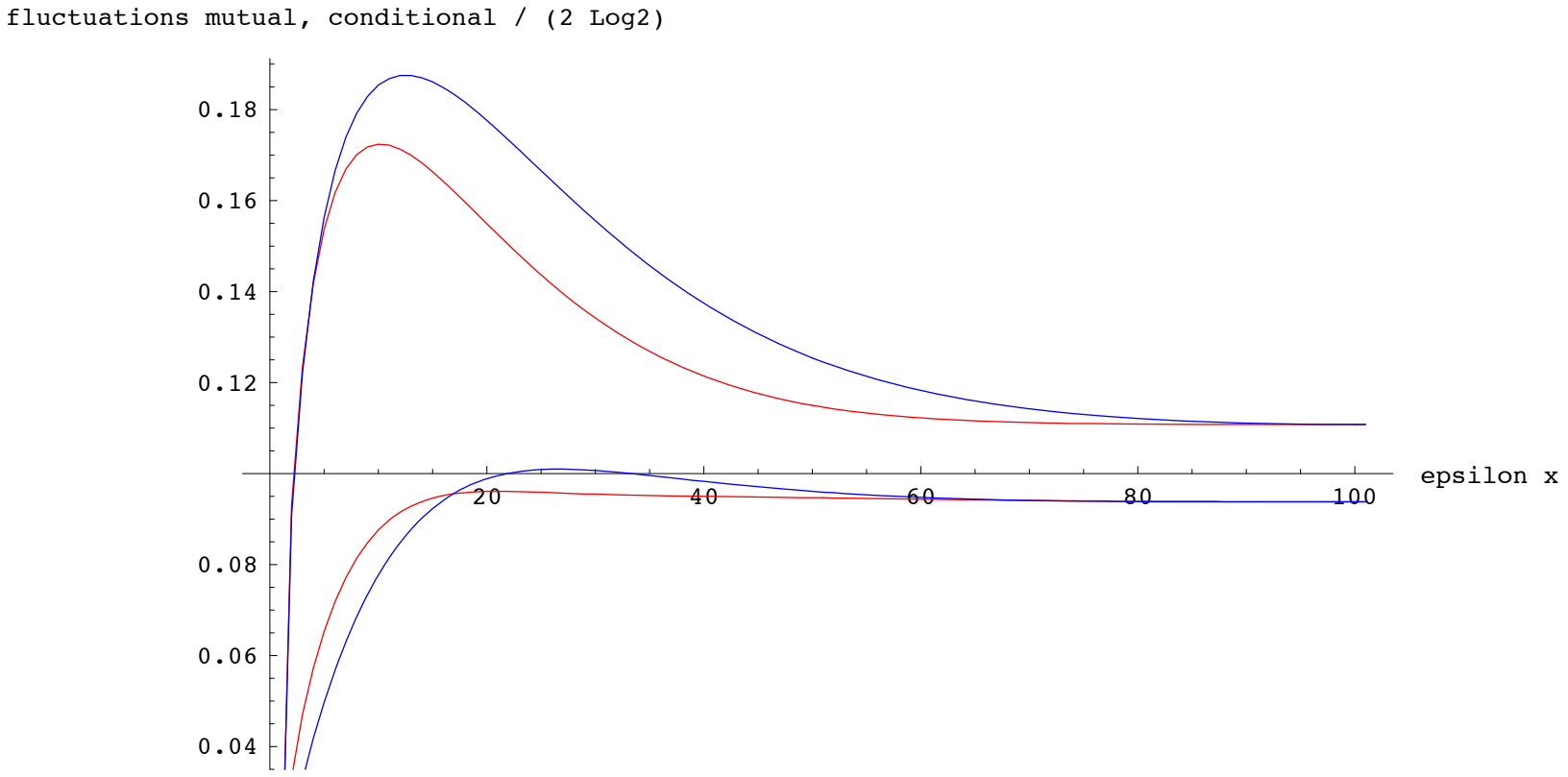
fluctuations mutual, conditional / (2 Log2)
0.18
0.16
0.14
0.12
0.08
0.06
0.04
20
40
60
80
100
epsilon x

FIG. 7. Fluctuations of M^{ij} (lower pair of curves) and $\Delta(j|i)$ (upper pair). For the upper pair, $\Delta[\Delta(j|i)]_\tau > \Delta[\Delta(j|i)]_s$. For the lower pair, the star fluctuation is larger than the tree one if $\varepsilon \lesssim 0.04$. It becomes smaller if $\varepsilon \gtrsim 0.04$.

We find that whether $\varepsilon \lesssim 0.04$ or $\varepsilon \gtrsim 0.04$, the fluctuation of the mutual information is (slightly) larger or smaller, respectively, for the star than for the tree. More important, it is seen that ΔM^{ij} tends to be somewhat smaller than $\Delta[\Delta(j|i)]$, which would point to the mutual information as a better criterion. Returning to Fig.5, however, where $< \Delta(j|i) >$ is significantly larger than the spurious non vanishing $< M^{ij} >$, it seems safer to stick to $\Delta(j|i)$ as a criterion for true correlations. Indeed, with such a likely smaller *relative* error, drops of $\Delta(j|i)$, as observed in the columns of the matrices of Sec.3, make a cleaner signal.

To summarize this Section, Sec.5, there is some evidence that *i)* a tabulation of noise levels is reasonably easy from elementary models and *ii)* in any case the "drop of $\Delta(j|i)$" criterion is always useful. The next Section, Sec.6, attempts to generalize such optimistic conclusions.

6. Generalizations to models with any number of bases and/or aminoacids. The inversion problem

In Section 2 we defined observables valid for any number S of "spin" values, while only the case $S = 2$ was investigated in Sections 3-5. For DNA/RNA, $S = 4$ and for proteins, $S = 20$. Numerical simulations for such cases, not reported here, do not give results which contradict, or differ significantly from, those discussed and listed in Sections 3-5. Similar decreases of weighted conditional entropy, for instance, are observed in case of influence between sites. Similar background "noise", due to fluctuations, are also present. There is a difference in the formalism to be used, however, pertaining to the coding of states and mutations. With $S = 4$ for instance, and a labeling of adenine, thymine, guanine and cytosine by 1, 2, 3 and 4, respectively, a mutation from adenine to guanine and a mutation from thymine to cytosine would be both coded by an increase of the spin label by 2. The problem of interpretation raised by such an ambiguity is hardly acceptable. There is thus less interest in generalizing Eqs.(12-21), unless each of the 16 possible mutations has a specific coding compatible with such generalized equations. The mutual information and the conditional entropy, nevertheless, retain their definitions without any difficulty in this formalism.

For $S > 2$ it is simpler to code the situation at site i by S occupation numbers n_s^i, $s = 1, 2...S$, restricted to two values, $n_s^i = 0$ and $n_s^i = 1$. Also an obvious constraint $\sum_s n_s^i = 1$ shall restrict the 2^S possibilities, offered by such a coding, to the only S meaningful ones. Except for such a constraint, the numbers n_s^i are otherwise independent random variables. Similar sets of occupation numbers n_t^j, with similar constraints $\sum_s n_s^j = 1$, will describe the situation at all the other sites j. The occurence numbers considered in Section 2 are then nothing but $N_{st}^{ij} \equiv \sum_{a=1}^{\mathcal{M}} n_s^i(a)\, n_t^j(a)$, where a labels each individual in the sampled population, and, obviously, $n_s^i(a)$ and $n_t^j(a)$ describe the presence or absence of residues s and t at sites i and j, respectively, in this individual a. Thus Eqs.(1-7) generalize, trivially. It is easy to create null models, with or without dynamical correlations between sites, generalizing the models and results of Sections 3-5. In particular there is no difficulty in generalizing Eqs.(23).

Rather than a tabulation and calibration of true and spurious amounts of correlations derived from such elementary models, the present Section, Sec.6, presents a solution of the following "inverse problem" [9]: **given** average values of observables, obtained from sampling,

$$\nu_s^i =< n_s^i >= \mathcal{M}^{-1} \sum_{a=1}^{\mathcal{M}} n_s^i(a), \qquad \nu_{st}^{ij} =< n_s^i n_t^j >= \mathcal{M}^{-1} \sum_{a=1}^{\mathcal{M}} n_s^i(a)\, n_t^j(a)\,, \tag{24}$$

and the additional constraints,

$$\sum_{s=1}^{S} n_s^i = 1, \quad i = 1, ...N\ , \tag{25}$$

what are the sites i and j whose contacts are compatible with such constraints?

The traditional probability distribution with maximum entropy for sequences a coded by degrees of freedom $\{n\} \equiv \{n_s^i\,, i = 1...N,\ s = 1...S\}$, under the constraints listed by Eqs.(24-25), reads,

$$P(a) = Z^{-1} \exp\left[-\sum_{i=1}^{N}\sum_{s=1}^{S} \lambda_s^i\, n_s^i(a) - \sum_{i=1}^{N}\sum_{j>i}^{N}\sum_{s=1}^{S}\sum_{t=1}^{S} \lambda_{st}^{ij}\, n_s^i(a)\, n_t^j(a) - \sum_{i=1}^{N} \lambda^i \left(1 - \sum_{s=1}^{S} n_s^i(a)\right)^{\nu} \right]\ , \tag{26}$$

where $\nu = 1$ and the partition function Z is a sum over all the 2^N sequences which can be constructed when the occupation numbers take on values 0 and 1,

$$Z = \sum_{\{n\}} \exp\left[-\sum_{i=1}^{N}\sum_{s=1}^{S} \lambda_s^i\, n_s^i - \sum_{i=1}^{N}\sum_{j>i}^{N}\sum_{s=1}^{S}\sum_{t=1}^{S} \lambda_{st}^{ij}\, n_s^i\, n_t^j - \sum_{i=1}^{N} \lambda^i \left(1 - \sum_{s=1}^{S} n_s^i\right)^{\nu} \right]\ . \tag{27}$$

The Lagrange multipliers λ_s^i and λ_{st}^{ij} are adjusted later in such a way as to satisfy the constraints, Eqs.(24),

$$\nu_s^i \ = \ -Z^{-1}\, \frac{\partial Z}{\partial \lambda_s^i} = Z^{-1} \sum_{\{n\}} n_s^i\, \exp\left[-\sum_{i=1}^{N}\sum_{s=1}^{S} \lambda_s^i\, n_s^i - \sum_{i=1}^{N}\sum_{j>i}^{N}\sum_{s=1}^{S}\sum_{t=1}^{S} \lambda_{st}^{ij}\, n_s^i\, n_t^j \right]\ , \tag{28a}$$

$$\nu_{st}^{ij} \ = \ -Z^{-1}\, \frac{\partial Z}{\partial \lambda_{st}^{ij}} = Z^{-1} \sum_{\{n\}} n_s^i\, n_t^j\, \exp\left[-\sum_{i=1}^{N}\sum_{s=1}^{S} \lambda_s^i\, n_s^i - \sum_{i=1}^{N}\sum_{j>i}^{N}\sum_{s=1}^{S}\sum_{t=1}^{S} \lambda_{st}^{ij}\, n_s^i\, n_t^j \right]\ . \tag{28b}$$

The remaining Lagrange multipliers λ^i are adjusted in such a way as to satisfy Eqs.(25), naturally, with $\nu = 1$. But nothing prevents us from taking *a priori* a unique and large positive value Λ for such remaining parameters while setting $\nu = 2$ in order to better enforce the constraints, Eqs.(25). The summations then run, in practice, over the only S^N admissible configurations, where one solves for e.g. the twentieth occupation number in terms of the other nineteen. Once Z is calculated via such suitable S^N configurations, this amounts, in the space of Lagrange multipliers, to solve for the minimum of the "free energy",

$$F = -\mathrm{Log}Z - \sum_{i=1}^{N}\sum_{s=1}^{S} \lambda_s^i\, \nu_s^i - \sum_{i=1}^{N}\sum_{j>i}^{N}\sum_{s=1}^{S}\sum_{t=1}^{S} \lambda_{st}^{ij}\, \nu_{st}^{ij}\,. \tag{29}$$

Convexity properties make this minimum unique [14]. The process therefore returns a unique set of parameters λ_s^i and λ_{st}^{ij}.

Define a "contact index" C^{ij} which vanishes if sites i and j do not interact. Conversely, define $C^{ij} = 1$ when such sites are close enough to induce interactions. It is reasonable to assign $C^{ij} = 0$ to those pairs ij of sites for which *all* the λ_{st}^{ij}, $s = 1...S$, $t = 1...S$, as obtained from the procedure which has just been described, are vanishing or small in some sense. Conversely, it is reasonable to assign $C^{ij} = 1$ when *at least one* of these λ_{st}^{ij} is large. This raises a problem of scale for the various λ_{st}^{ij} 's. We shall assume that such numbers, or rather their absolute values, cluster into two groups, namely the "small" and the "large" $\left|\lambda_{st}^{ij}\right|$'s, respectively. For those pairs $\{ij\}$ for which *every* $\left|\lambda_{st}^{ij}\right|$ is small, it can be concluded that $C^{ij} = 0$. Conversely, for those pairs of sites for which *at least* one of the $\left|\lambda_{st}^{ij}\right|$'s is a member of the other cluster, namely this $\left|\lambda_{st}^{ij}\right|$ is interpreted as "large", it can be concluded that $C^{ij} = 1$.

Preliminary calculations [9] show that this procedure may eliminate spurious chainings. That such a result is possible, while not mandatory, is easy to understand from Eq.(26), which, although parametrized by one- and two body features only, trivially allows for observables of any higher rank. For instance, it is straightforward to calculate three body observables such as, e.g., $< n_s^i \, n_t^j \, n_u^k >$.

To summarize this Section, Sec.6, multiple valued spin models are available, and simple enough, to study the influence of statistical fluctuations upon remote site correlations. Beyond numerical tabulations of various noise levels, and corresponding confidence levels, for M^{ij} and/or $\Delta(j|i)$ for trees and stars, the maximum entropy procedure described by Eqs.(26-28) (see also [9]) provides convenient estimates of links λ_{st}^{ij} between sites, and of resulting contact indices C^{ij}.

7. Discussion and conclusions

In the search for evidences of contacts between seemingly remote sites of biological sequences, this paper essentially reports three results and one failure.

The first result is the validation of a criterion related to algorithmic information theory, incorporating conditional probablities. The indicator $\Delta(j|i)$ defined by Eqs.(5-6) drops significantly when there is a causal relation between sites i and j, and the drop does not seem to be overly sensitive to statistical fluctuations or to correlations induced by shared ancestry. The more familiar mutual information between sites, M^{ij}, does not seem to be so robust.

The failure is related to this first result. We did not find a clear signature for nonreciprocal influences. From a physical point of view this is not a major issue, since action and reaction are reciprocal. In the case of historical evolutions, though, with delayed actions, this problem is not without interest and deserves further investigation.

The second result is the large available class of *practicable* "spin" models, null models without interactions or more realistic ones with intersite influences. As we discussed in some detail, both analytically and numerically, such models are quite useful to understand the rôle of statistical fluctuations linked to finite sampling and those associated with effects of shared ancestry of sequences. The point is, naturally, that various degrees of freedom of the problem are not

independent variables and the central limit theorem is violated. In the comparison between models of evolution down a tree versus star topologies, an intuitive "uncertainty principle" was formalized : those evolutions which favor similarity between individuals amplify collective deviations from ancestral properties. It is trivial to generalize to multivalued spins the "sign argument" used after Eq.(20). Namely, any positive result for a correlation $< n_s^i(a) n_s^i(b) > - < n_s^i >^2$ between individuals will induce a lowering of the interindividual $< \left[n_s^i(a) - n_s^i(b) \right]^2 >$ dispersion and, simultaneouly, an increase of the fluctuation of the "center of mass" average $\mathcal{M}^{-1} \sum_a n_s^i(a)$ over the population. Above all, such models are valuable for a numerical tabulation of such fluctuations. Any signal exceeding the fluctuations deduced from the models thus exceeds a confidence threshold and can be taken as a reliable evidence.

The third result is the maximum entropy solution of the inverse problem "given correlations, find the couplings between sites". There is a risk that the couplings may not cluster into two groups of, respectively, large and small couplings. But the convexity of this algorithm and the unicity of the couplings provided by this solution are worth consideration. In addition to the preliminary results presented in [9], and the results reported here, a systematic investigation of the application of this formalism to real biological sequences [12] is in progress.

Acknowledgments: The authors thank the Santa Fe Institute where part of this work was performed. The research of Lapedes and Liu was financially supported by the U.S. Department of Energy. Lapedes thanks Gary Stormo for numerous helpful conversations concerning aspects of this work.

---

[1] Korber, B., Farber, R., Wolpert, D., Lapedes, A. *Covariation of Mutations in the V3 Loop of HIV-1: An Information Theoretic Analysis* Proc. Nat. Acad. Sciences, Vol 90, 7176 (1993)
[2] Gutell, R.R., Power, A., Hertz, G.Z., Putz, E. and Stormo, G.D., Nucl. Acids Res. 20:5785-5795
[3] Gobel, U., Sander, C., Schneider, R., Valencia, A. *Correlated Mutations and Residue Contacts In Proteins* Proteins: Structure, Function, Genetics, vol 18:309-317
[4] Clarke, N. *Covariation of Residues in the Homeodomain Sequence Family* Protein Science 4:2269-2278
[5] Shindyalov, I., Kolchanov, N., Sander, C. *Can Three Dimensional Contacts in Protein Structures be Predicted by Analysis of Correlated Mutations?* Protein Engineering 7:349-358
[6] Neher, E., *How Frequent are Correlated Changes in Families of Protein Sequences?* Proc. Natl. Acad. Sci. USA 91:98-102
[7] Taylor, W., Hatrick, K. *Compensating Changes in Protein Multiple Sequence Alignments* Protein Engineering 7:341-348
[8] Thomas, D., Casari, G., Sander C. *The Prediction of Protein Contacts From Multiple Sequence Alignments* Protein Engineering 9:941-948
[9] Lapedes, A.S., Giraud, B.G., Liu, L.C. and Stormo, G.D., *Correlated Mutations In Protein Sequences: Phylogenetic and Structural Effects*, to be published in Proceedings of the AMS/SIAM Conference "Statistics in Molecular Biology" (Seattle, WA 1997), also Santa Fe Institute Working Paper number 97-12-088
[10] Stanley, H. *Introduction to Phase Transitions and Critical Phenomena* The International Series of Monographs on Physics, Oxford University Press Inc., Oxford and New York
[11] Cover, T., Thomas, J. *Elements of Information Theory* Wiley Series in Telecommunications, John Wiley and Sons
[12] Lapedes A.S., Giraud B.G., Liu L.C., Stormo, G.D., *Disentangling Chains of Correlated Mutations In Protein Sequences: Approaches to Predicting Contacts*, preprint
[13] Zurek, W. *Thermodynamic Cost of Computation, Algorithmic Complexity and the Information Metric* Nature 341, p. 119 (1989)
[14] Alhassid Y., Agmon N., Levine R. *An Upper Bound for The Entropy annd its Application to the Maximal Entropy Problem* Chem. Phys. Letter vol. 53, p. 22 (1978)