



An index of substitution saturation and its application

Xuhua Xia, *et al*

Erik Barry Erhardt, April 27, 2005

[Introduction, Motivation](#)

[Methods](#)

[Results](#)

[Application](#)

[Conclusions](#)

[Home Page](#)

[Title Page](#)



Page 1 of 21

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

Substitution Saturation

Outline

1. Introduction.
2. Methods.
3. Results.
4. Application.
5. Conclusions.



Introduction, Motivation

Methods

Results

Application

Conclusions

Home Page

Title Page



Page 1 of 21

Go Back

Full Screen

Close

Quit

1. Introduction

1.1. Phylogenetic reliability

Five problems:

1. Reliability of sequence alignment.
2. Substitution rates vary substantially over sites.
3. Nucleotide frequencies change.
4. Long-branch attraction.
5. Lost phylogenetic information due to substitution saturation. ★

1.2. Substitution saturation

- Problem for phylogenetic analysis involving deep branches.
- Full saturation: depend entirely on similarity in [essentially random] nucleotide frequencies.
- So conservative genes often used.

1.3. Codons

- Protein genes consist of codons.
- Each codon consists of 3 nucleotides, giving $4^3 = 64$ possible codons, determining 20 amino acids.
- Generally, the first two codons determine the amino acid and the third is free to vary.
- Third codon position is the most variable.
- Second codon the most conservative.
- Third codon is often used to help estimate divergence time.
- However if experienced substitution saturation, may contain no phylogenetic information.

1.4. Does molecular sequence contain phylogenetic information?

- Present an entropy-based index of substitution saturation.
- Statistically test whether saturation has occurred.

2. Methods

2.1. Concepts

- Suppose N aligned sequences with L nucleotides each, with nucleotide frequencies $P_A, P_C, P_G,$ and P_T .
- Consider no substitution, then nucleotides will be identical for at each site for all sequences.
 - If all As, $P_A = 1, P_C = P_G = P_T = 0$.
- In terms of information theory, the entropy at site i is
$$H_i = - \sum_{j=1}^4 p_j \log_2 p_j. \quad ^1$$
- With no substitutions, $H_i = 0$, and H_i increases to 2 when frequencies are all equal at $\frac{1}{4}$.
- Sample means and variances of H are easily calculated over all L sites.

¹Claude Shannon was interested in juggling, unicycling, and chess. He also invented many devices, including a chess-playing machine, a rocket-powered pogo stick, a wearable computer to predict the result of playing roulette, and a flame-throwing trumpet for a science exhibition.

2.2. Sample Statistics

- Sample mean and variance.

$$\bar{H} = L^{-1} \sum_{i=1}^L H_i$$

$$\text{Var}(H) = (L - 1)^{-1} \sum_{i=1}^L (H_i - \bar{H})^2$$

[Introduction, Motivation](#)

[Methods](#)

[Results](#)

[Application](#)

[Conclusions](#)

[Home Page](#)

[Title Page](#)



[Page 5 of 21](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

2.3. Expected Values

- Full Substitution Saturation (FSS).
- Expected values based on multinomial distribution.

$$\begin{aligned}
 H_{\text{FSS}} &= \sum_{N_A=0}^N \sum_{N_C=0}^N \sum_{N_G=0}^N \sum_{N_T=0}^N \frac{N!}{N_A!N_C!N_G!N_T!} \\
 &\quad \times P_A^{N_A} P_C^{N_C} P_G^{N_G} P_T^{N_T} \left(- \sum_{j=1}^4 p_j \log_2 p_j \right) \\
 \text{Var}(H_{\text{FSS}}) &= \sum_{N_A=0}^N \sum_{N_C=0}^N \sum_{N_G=0}^N \sum_{N_T=0}^N \frac{N!}{N_A!N_C!N_G!N_T!} \\
 &\quad \times P_A^{N_A} P_C^{N_C} P_G^{N_G} P_T^{N_T} \left(\sum_{j=1}^4 p_j \log_2 p_j - H_{\text{FSS}} \right)^2
 \end{aligned}$$

- $N = N_A + N_C + N_G + N_T$, $p_j = N_i/N$

2.4. Test of Substitution Saturation

- Test whether observed \bar{H} is significantly smaller than H_{FSS} .
- Index of substitution saturation, $I_{\text{SS}} = \bar{H}/H_{\text{FSS}}$.
- Clearly, sequences have experienced severe substitution saturation when I_{SS} approaches 1.
- But, sequences fail to recover the true phylogeny long before the full substitution saturation is reached.
- So, calculate a critical value $I_{\text{SS,C}}$ for a set of sequences with known properties.
- If $I_{\text{SS}} > I_{\text{SS,C}}$ we will conclude that severe substitution saturation has occurred, and these sequences should not be used to construct phylogenetic topologies.
- $I_{\text{SS,C}}$ can be studied through simulation of an experimental set of topologies, number Operational Taxonomic Units (OTUs, or N_{OUT}), sequence length (SeqLen), nucleotide frequencies, and transition/transversion ratio.

2.5. Computer Simulation

- PAML/EVOLVER for evolutionary simulation according to F84.
- The α/β ratio varied from 1 to 10.
- The nucleotide frequencies of the four nucleotides varied from 0.1 to 0.9, subject to the constraints that the summation equals 1.
- Effect of transition/transversion ratio and nucleotide frequencies on $I_{SS.C}$ is negligible compared to the effect of topology, N_{OTU} , and SeqLen.

[Home Page](#)

[Title Page](#)



Page 8 of 21

[Go Back](#)

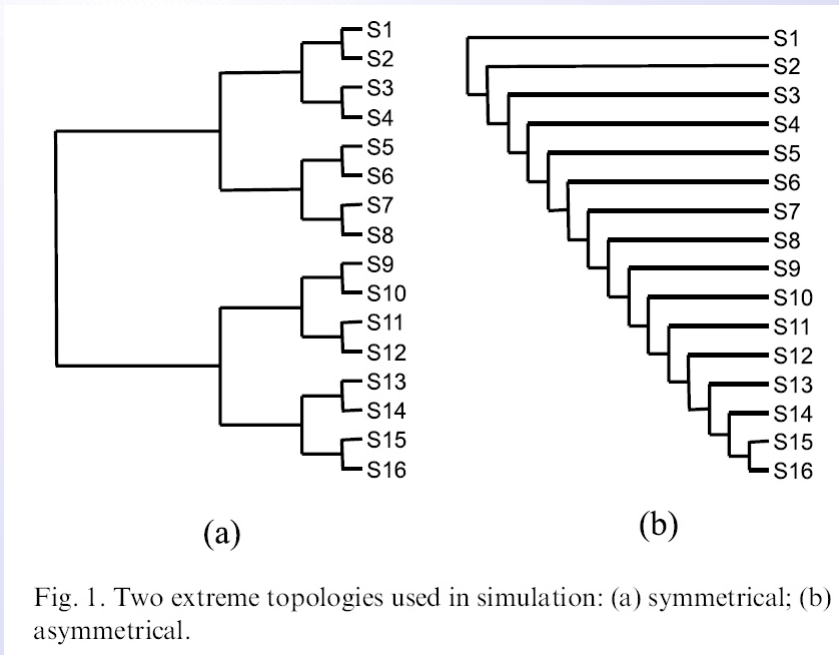
[Full Screen](#)

[Close](#)

[Quit](#)

2.6. Extreme Topologies (Fig. 1)

- Consider best- and worst-case topologies in simulation.



2.7. Factor Combinations

- The N_{OTU} values are 4, 8, 12, 16, 20, 24, 28, and 32.
- When N_{OTU} values are 12, 20, 24, and 28, there is no perfectly symmetrical topology as in Fig. 1a, and multiple quasi-symmetrical topologies were used.
- For example, when $N_{\text{OTU}} = 12$, then we obtain multiple topologies by randomly pruning of a four-OTU symmetrical subtree from the symmetrical 16-OTU topology.
- SeqLen values 500, 1500, 2500, 3500, 4500, and 5500.
- Longer sequences should alleviate effect of substitution saturation as long as sequences have not experienced full substitution saturation.
- $I_{\text{SS.C}}$ value should be greater with a set of long sequences than with a set of short sequences, everything else being equal.
- Tree length varies from 1 to 29 for the symmetrical topology and from 1 to 19 for the asymmetrical topology (1, 3, 5, ...).



Introduction, Motivation

Methods

Results

Application

Conclusions

Home Page

Title Page



Page 11 of 21

Go Back

Full Screen

Close

Quit

- For a given topology and N_{OTU} , the longer the tree length, the greater the substitution saturation and the greater the I_{SS} value.
- Which I_{SS} value the sequences will be too substitutionally saturated to recover the true tree?
- This particular I_{SS} value is taken as the $I_{\text{SS,C}}$ value.
- By doing a large number of simulations, we can determine $I_{\text{SS,C}}$ empirically for a given SeqLen, a given N_{OTU} , and a given topology.

2.8. Methods

- Trees with tree length shorter than 1 not used since too few substitutions to recover true tree.
- Each topology simulated 100 times.
- Phylogenetic reconstruction to find the proportion of trees correctly reconstructed P_{true} .
- The neighbor-joining (NJ) and maximum likelihood (ML) method with F84 models yield essentially the same P_{true} values.
- NJ results are presented.
- Data Application: Regier and Shultz (1997) 16 sequences of the EF-1 α gene from major arthropod groups and putative outgroups.
- Aligned by first translating into amino acid sequences, aligned, and the nucleotide sequences were aligned against aligned amino acid sequences by using DAMBE.

3. Results

3.1. Simulation Studies (Fig. 2)

- Ability in recovering the true tree decreases with the total tree length (i.e., the degree of substitution saturation).
- Effect of substitution saturation is alleviated by increasing SeqLen.
- $I_{SS,C}$ is the value corresponding to the critical tree length (TL_C) which is when P_{true} is 95% of the maximum P_{true} value.
- Often no tree length at which the true tree is recovered 100%.
- P_{true} value decreases when the tree length (TL) approaches zero implying the rarity of substitution saturation (not shown).

Introduction, Motivation

Methods

Results

Application

Conclusions

Home Page

Title Page



Page 13 of 21

Go Back

Full Screen

Close

Quit

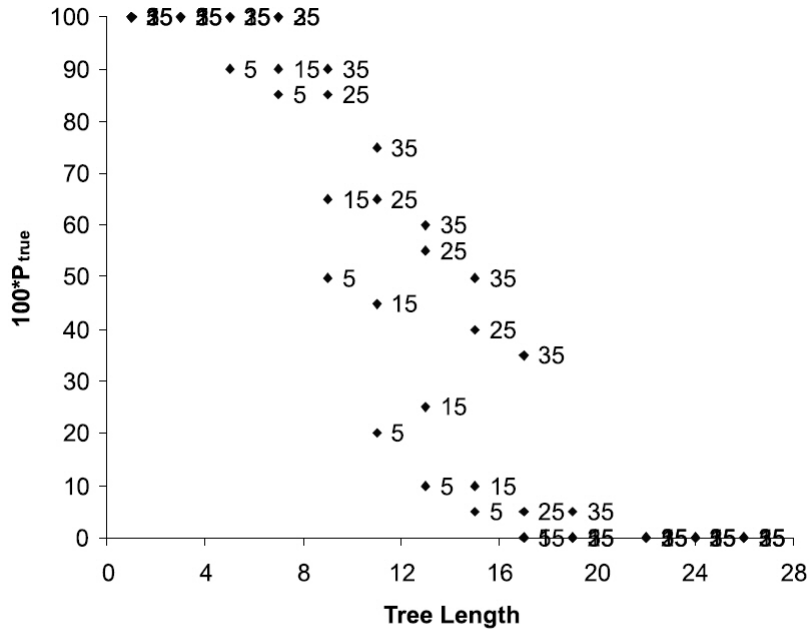


Fig. 2. The proportion of true trees found (P_{true}) depends on the tree length and the sequence length. The data shown are for $N_{\text{OTU}} = 8$. The pattern remains the same with different number of OTUs. Data labels indicate the sequence length in 100, i.e., a number of 5 means 500 bp.

3.2. Critical Index of Substitution Saturation, $I_{SS.C}$ (Fig. 3)

- $I_{SS.C}$ value depends on SeqLen, topology, and N_{OTU} in the tree.
- For given SeqLen, $I_{SS.C}$ decreases with increasing N_{OTU} .
- This decrease is more severe for asymmetrical topology.
- Asymmetrical tree more susceptible to substitution saturation.
- If OTUs likely to be phylogenetically related by asymmetrical topology, should increase the sequence length.
- $I_{SS.C}$ values increase with SeqLen, increasing SeqLen can alleviate the problem of substitution saturation.
- However, increase of $I_{SS.C}$ levels off beyond 4000 bp.
- For recovering deep phylogenies, better to use short conserved sequences than long highly variable sequences (or gene order).
- Note $I_{SS.C}$ small for $N_{OTU} = 12, 20, 24, 28$ since these N_{OTU} values cannot be perfectly symmetrical.
- Even slight deviation from perfect symmetry can decrease $I_{SS.C}$.

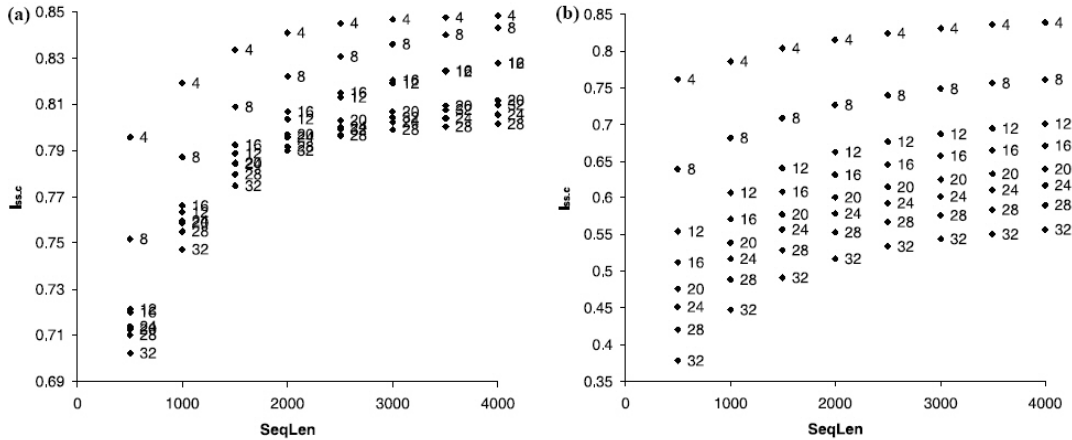


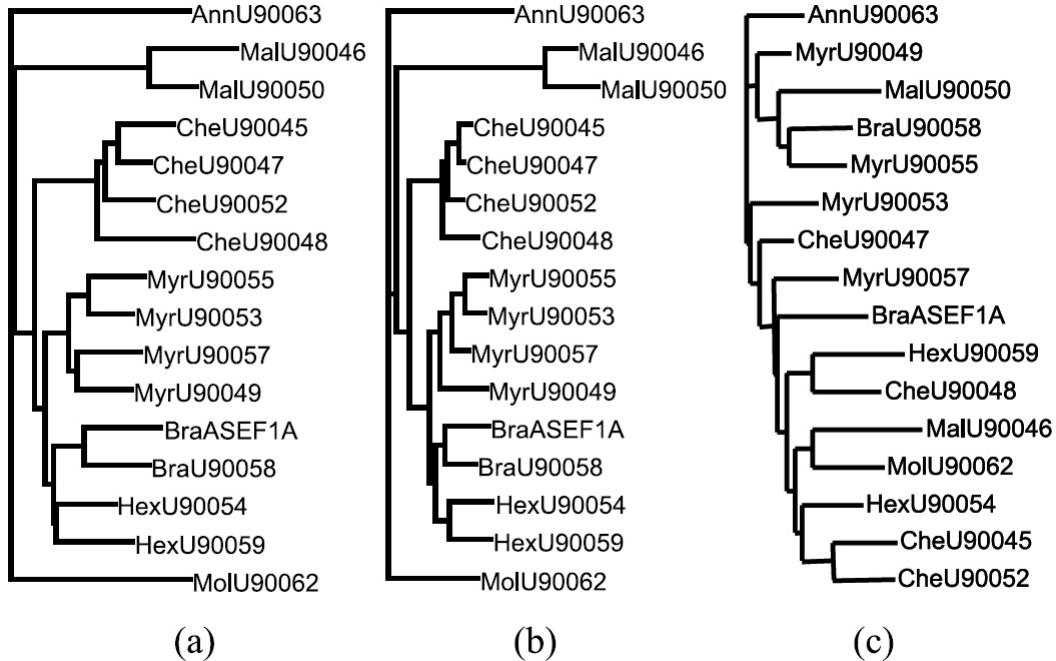
Fig. 3. The critical index of substitution saturation (I_{sac}) depends on the sequence length (SeqLen) and the number of OTUs (N_{OTU}). Data labels are N_{OTU} values: (a) with a symmetrical topology; (b) with an asymmetrical topology.

4. Application

4.1. Application of the method to real sequences

- First, second, and third codon positions of EF-1 α sequences have I_{SS} values 0.2093, 0.1115, and 0.6636.
- The $I_{SS,C}$ value, given $N_{OTU} = 16$ and SeqLen=350, is 0.7026 (symmetrical) and 0.4890 (asymmetrical).
- I_{SS} is much less than $I_{SS,C}$ at the first and second codon positions.
- So little evidence for substitution saturation at these positions.
- Third codon position $I_{SS} = 0.6636$ is less than 0.7026 (symmetrical) but larger than 0.4890 (asymmetrical).
- So evidence that third codon position has experienced so much substitution saturation that it is only marginally useful when the true tree is symmetrical and useless if the true tree is asymmetrical for reconstructing topology.

- The resulting phylogenetic trees based solely on the first, second, and third codon positions are shown in Fig. 4a, b, and c.
- Reconstructed with third codon positions is poor.



4.2. Other Models

- Applicability of the test appears reasonable under both the rates-across-sites (RAS) model and the covarion hypothesis.



Introduction, Motivation

Methods

Results

Application

Conclusions

Home Page

Title Page



Page 19 of 21

Go Back

Full Screen

Close

Quit

5. Conclusions

- The entropy-based index can be used to test whether aligned sequences can be useful in phylogenetics.



Introduction, Motivation

Methods

Results

Application

Conclusions

Home Page

Title Page



Page 20 of 21

Go Back

Full Screen

Close

Quit



Introduction, Motivation

Methods

Results

Application

Conclusions

Home Page

Title Page



Page 21 of 21

Go Back

Full Screen

Close

Quit

$\text{\LaTeX} 2_{\epsilon}$ replaces all wordprocessors!

↷ This **document** was joyfully produced with $\text{\LaTeX} 2_{\epsilon}$ using the `pdfscreen.sty` package — and nothing else.

⊖ NO PowerPoint. † No Microsoft. ♡ No Problems.

◁ Just *beautiful*, **functional** documents with \LaTeX .

ΘΥΓ· Visit **TUG** for liberation from ugly and cumbersome giants.