UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

MOTIF FINDER IN DNA SEQUENCES USING EXPECTATION MAXIMIZATION (EM) ALGORITHM

Devin Vincent Tark (devinvt2)
Gowtham Kuntumalla (gowtham4)
Xiaowei Lyu (lv7)

CS 412 Spring 2020 - Project Report

1 INTRODUCTION

A ‘motif’ is a pattern in a sequence. For example, in DNA sequences (which are sequences over the alphabet A,C,G,T), an example of a motif is the pattern ‘TCACGTG’. A slightly more complex motif is the pattern TC[A/C]CGTG, which represents ‘either TCACGTG or TCCCGTG’. An occurrence of a motif in a given DNA sequence is called a ‘site’.

A more popular form of a motif is that of the ‘position weight matrix’ (PWM). This is a probabilistic pattern. An example of a PWM (with motif length = 5, sequence alphabet = A,G,C,T- length = 4) is shown below:

Motif (PWM)

$$\begin{array}{cccc} A & G & C & T \\ \end{array}$$
$$\begin{bmatrix} 0.5 & 0.1 & 0.25 & 0.15 \\ 0.3 & 0.3 & 0.1 & 0.3 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.2 & 0.2 & 0.2 & 0.4 \\ 0.1 & 0.4 & 0.4 & 0.1 \end{bmatrix} \tag{1.1}$$

The ‘information content’ of a PWM ‘W’ of length L with alphabet length = 4 is defined by:

$$ICPC = \sum_{i=1}^{L} \sum_{\alpha \in \{A,G,C,T\}} W_{i\alpha} log_2(\frac{W_{i\alpha}}{0.25}) \tag{1.2}$$

The information content represents how ‘sharp’ the pattern is. For example, if every position is uniformly distributed among the 4 characters, the information content is 0. If a position prescribes an ‘A’ with probability 1 and all other characters are disallowed (probability 0), that position contributes log(4) to the information content, the maximum possible contribution of a single position of the motif.

One set of popular algorithms that have been developed in the past 30 years focused on the Expectation Maximisation approach. [1], MEME algorithms [2], [3] and EXTREME algorithm [4] There is also a website which implements wide varieties of software similar to the MEME algorithm [5].

2 Problem Statement

The project involves developing a “motif finding” program and testing it. The goal of motif finding is the detection of unknown signals in a set of DNA sequences. In our project, given generated sequences, the program is going to find expectation of motifs and sites. The three major components of the implementation are:

- Building a benchmark
- Implementing the “motif finder”
- Evaluating the motif finder on the benchmark and making intelligent inferences.

We’ll use Expectation-Maximization (EM) as our “motif finder”. We choose EM since it is a course-related algorithm and is effective to find hidden variables behind large datasets.

3 Description of EM Algorithm

The Expectation-Maximization (EM) is a family of algorithms for learning probabilistic models in problems that involve latent variables (variables that we cannot directly observe but rather inferred from the observed variables). It can find maximum-likelihood estimates for hidden model parameters. It is an iterative way to approximate the maximum likelihood function. How the algorithm works can be described as below:

1. **Initialization**: Get an initial estimate for parameters θ.This can just be a random initialization or we can use EM based heuristic for choosing a better starting point.
2. **Expectation Step**: Assume the parameters from the previous step are fixed, compute the expected values of the latent variables (or more often a function of the expected values of the latent variables).

3. **Maximization Step**: Given the values you computed in the last step (essentially known values for the latent variables), estimate new values for θ that maximize a variant of the likelihood function.

4. **Exit Condition**: If likelihood of the observations have not changed much, exit; otherwise, go back to Expectation Step.

Bailey and Elkan [2] uses EM algorithm for discovering motifs in a group of related DNA or protein sequences. In that case, the latent variable is where the motif starts in each training sequence. The algorithm takes as input a group of DNA or protein sequences (the training set) and outputs expected motifs and sites as requested.

3.1 High level Pseudo Code

A high level view of this project algorithm is given below. Much of the algorithm has been adapted from the papers by Bailey and Elkan [2, 3].

Algorithm 1: Expectation-Maximization

```
Initialization: Find the most common subsequences as starting points;
Generate an initial PWM based on dataset;
while $iterations < n\_iter$ do
    Estimate motif positions $Z_{ij}$ from motif matrix (E-Step);
    Update the PWM from all positions $Z_{ij}$ (M-Step);
end
```

4 Experiments

We modified the number of iterations n_iter, which is used to mimic iteration convergence criterion, and hyper parameter $beta$, which is used in the calculation of letter frequency in the maximization step. Experiments (-computer simulations) were run using

$$n\_iter \in \{10,\ 100,\ 500,\ 1000\}$$

$$beta \in \{0.001,\ 0.01,\ 0.02,\ 0.025,\ 0.05,\ 0.1\}$$

Default values that are used in running experiments are $n\_iter = 100$ and $beta = 0.01$.
In the following figures, we show the overlapping positions, overlapping sites, relative entropy and runtime of different parameter sets. In each figure, point annotation represents experimental dataset parameters: (Information content per column (ICPC), Motif length, Sequence length, Number of sequences)

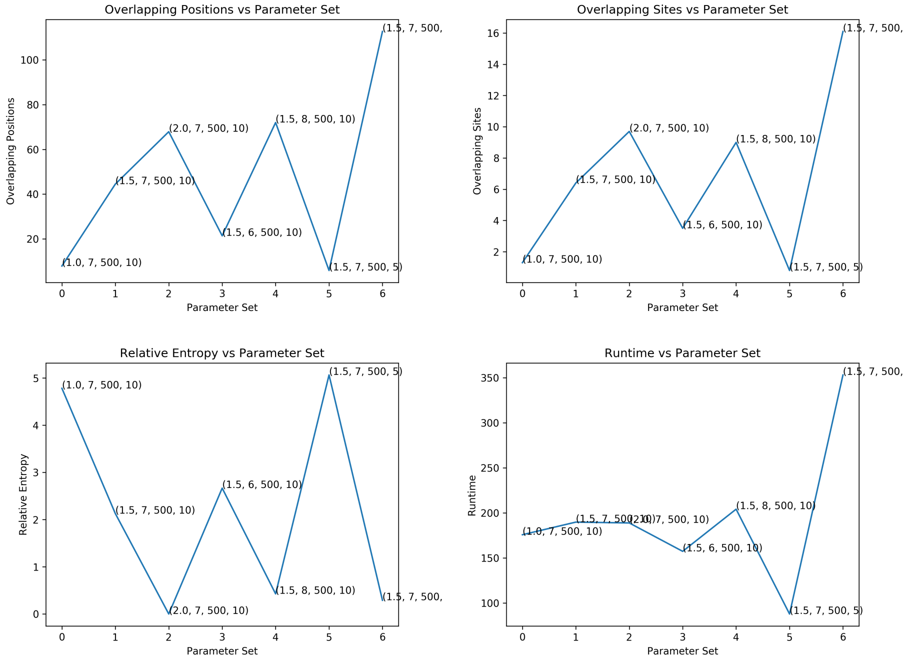

Figure 4.1: $n\_iter = 1000$, $beta\ =\ 0.01$

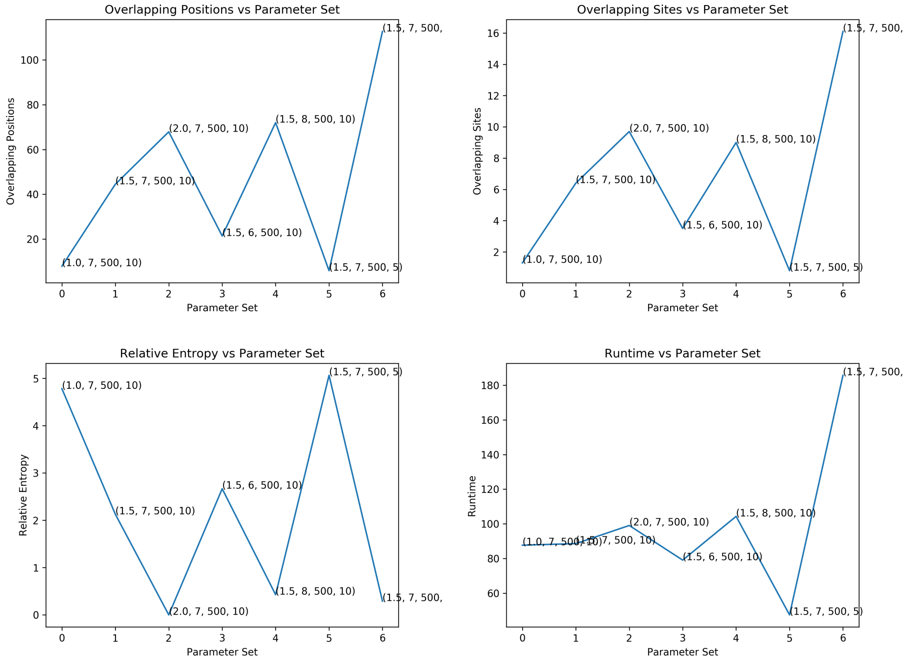

Figure 4.2: $n\_iter = 500$, $beta\ =\ 0.01$

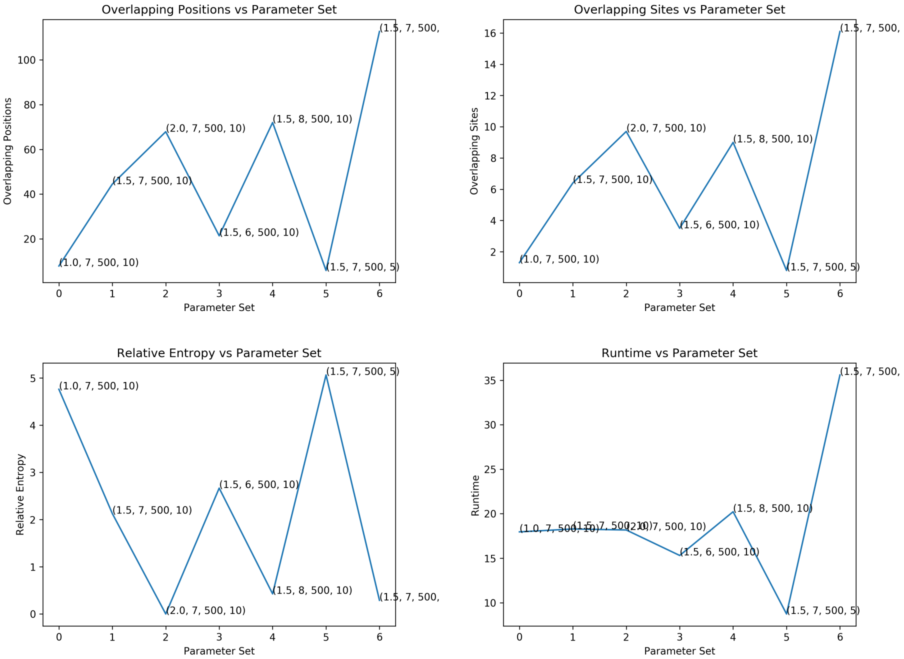

Figure 4.3: $n\_iter = 100$, $beta\ =\ 0.01$

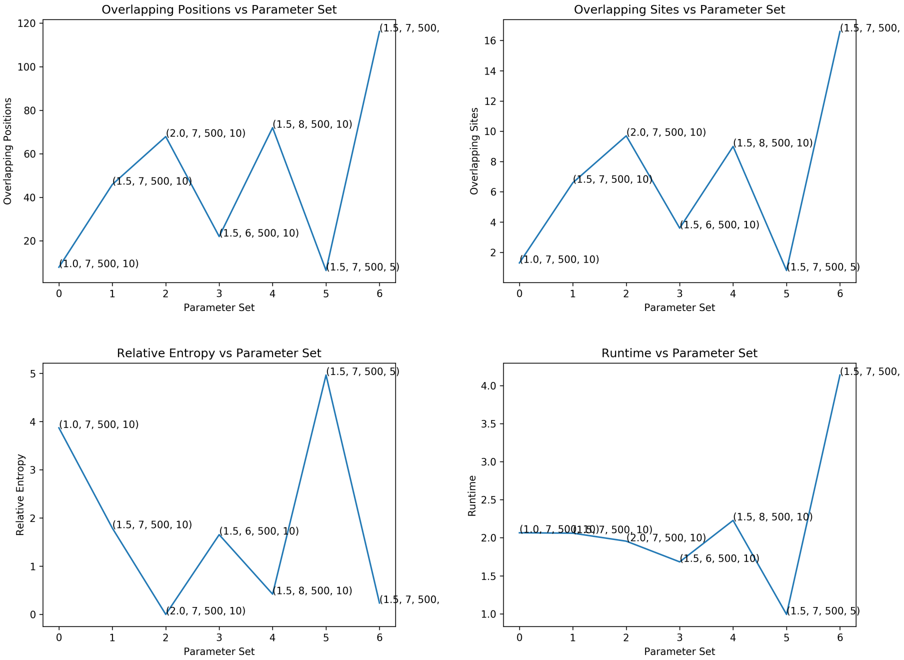

Figure 4.4: $n\_iter = 10$, $beta\ =\ 0.01$

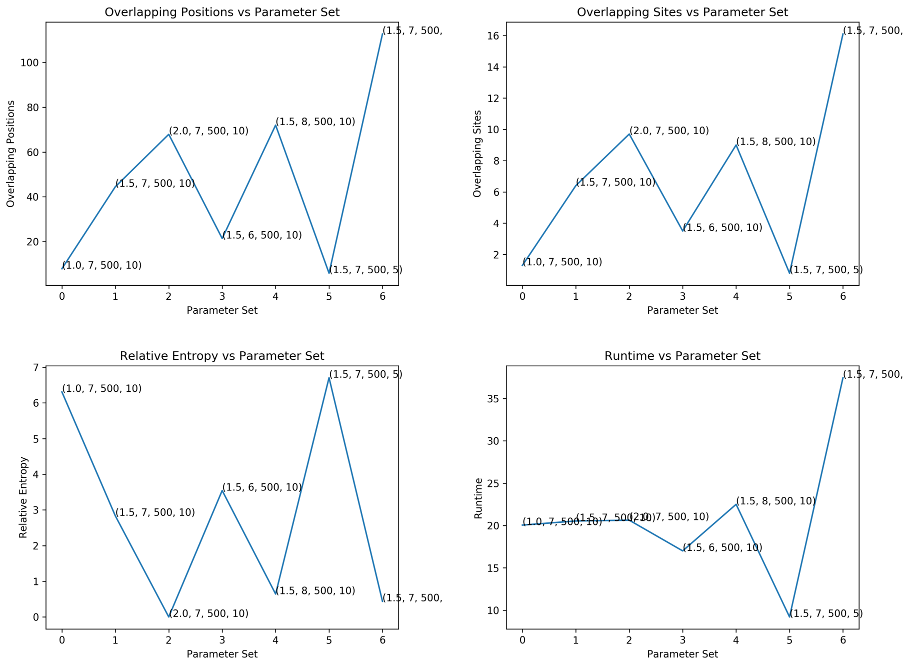

Figure 4.5: $n\_iter = 100$, $beta\ =\ 0.001$

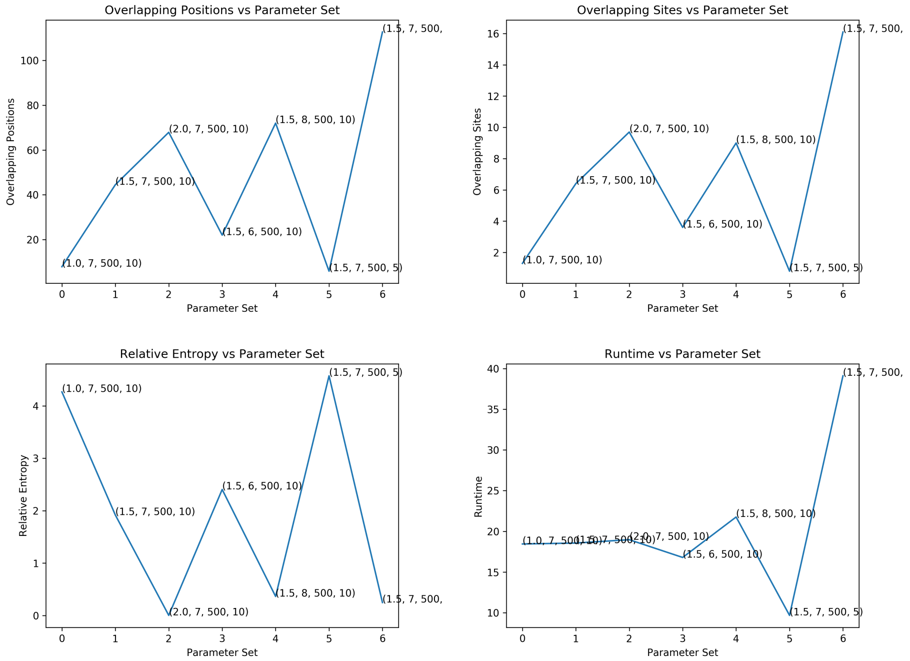

Figure 4.6: $n\_iter = 100$, $beta\ =\ 0.02$

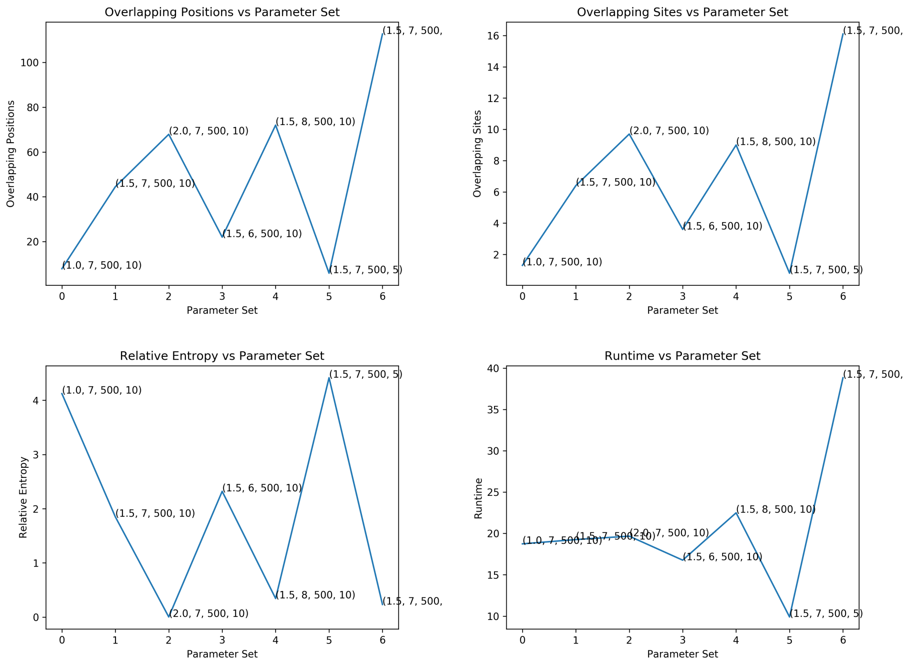

Figure 4.7: $n\_iter = 100$, $beta\ =\ 0.025$

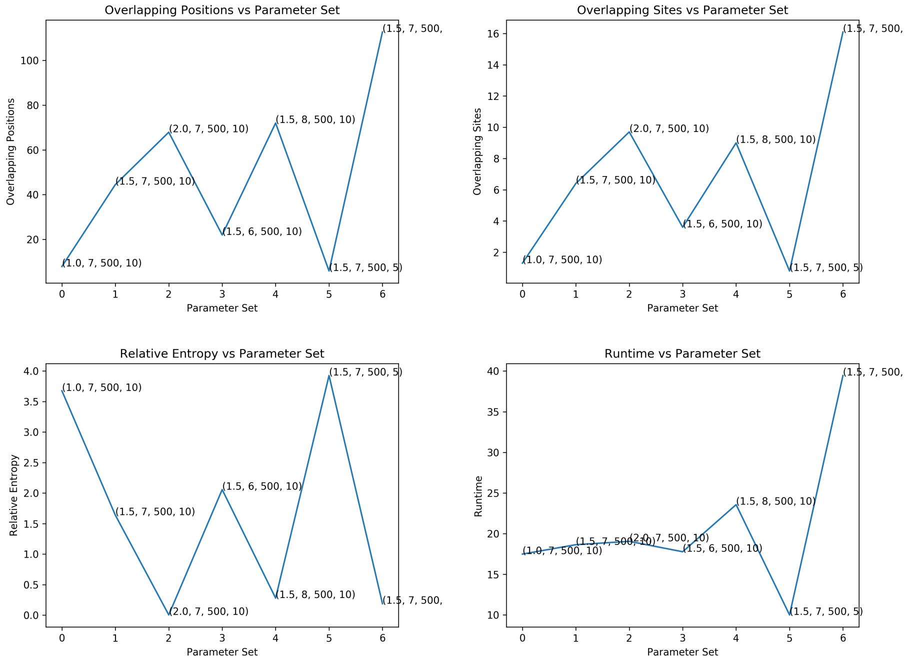

Figure 4.8: $n\_iter = 100$, $beta\ =\ 0.05$

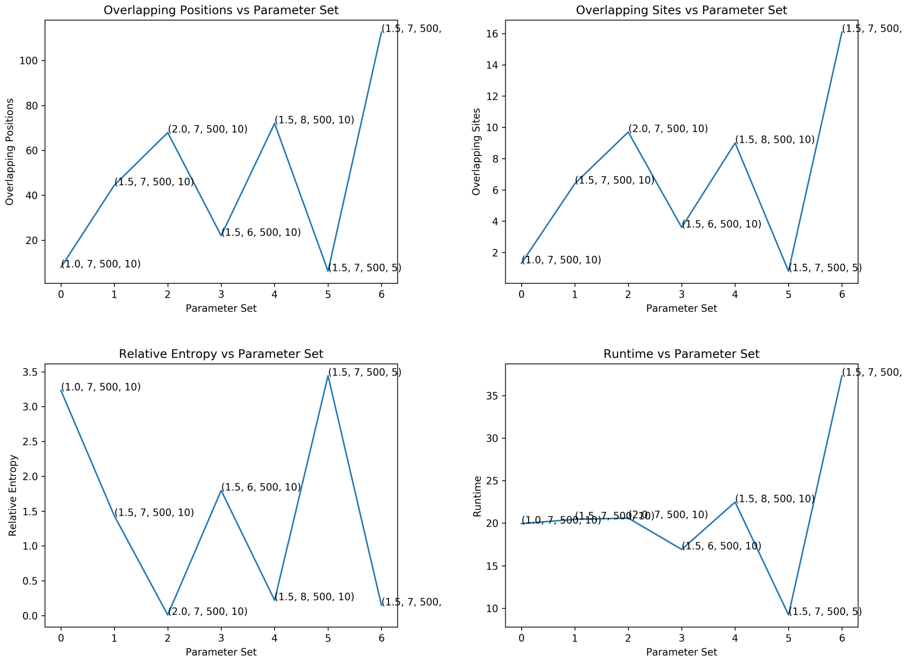

Figure 4.9: $n\_iter = 100$, $beta = 0.1$

5 Results & Conclusions

We find that

- Initialization is an important step and takes a significant role in finding an optimal solution
- Higher information content (ICPC) is associated with better predictions. For instance, this algorithm perfectly predicts the motif sites for ICPC = 2.
- Algorithm usually converges for for $n\_iter > 100$. In case of higher ICPC the convergence is even quicker.
- Changing hyper parameter $beta$ doesn't have a strong trend, but higher values seem to have lower relative_entropy. $beta = 0.1$ has the lowest relative_entropy amongst the tested values. Lower the relative entropy value, better is the accuracy of the algorithm.

6 Code Repository

Project codes and experiments are available at `https://github.com/gowthamkuntumalla/Motif_Finding_DNA`

REFERENCES

[1] C. Lawrence and A. Reilly, "An expectation maximization (em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences.," *Proteins*, 1995.

[2] T. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 1994.

[3] T. Bailey and C. Elkan, "Unsupervised learning of multiple motifs in biopolymers using expectation maximization," *Machine Learning*, 1995.

[4] D. Quang and X. Xie, "EXTREME: an online EM algorithm for motif discovery," *Bioinformatics*, vol. 30, pp. 1667–1673, 02 2014.

[5] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble, "MEME Suite: tools for motif discovery and searching," *Nucleic Acids Research*, vol. 37, pp. W202–W208, 05 2009.