Computing Microbial Interactions and Communications - COMIC

BioComputing and Computational Biology Group
Department of Computer Science
The University of Liverpool

3/7/02

Richard Gregory
Computer Science
The University of Liverpool

Ray Paton
Computer Science
The University of Liverpool

Jon Saunders
School of Biological Sciences
The University of Liverpool

Henry Wu
Electrical Engineering
The University of Liverpool

Abstract

For bacteria, evolution is the key the adaptability. The mechanism most often cited is via small mutations to the genetic code over many many generations. It has been shown that bacteria are subject to many more types of mutation than this simple small scale alteration, the question remains of how much bacteria gain from each mutation type and it is that question that this bacterial model aims to answer by simulating bacterial life at the genetic scale and implementing the variety of genetic adaption mechanisms. Described here are the genetics, the model and some results showing evolution towards the goal of learning to connect a food source gradient with movement toward the food source.

1 Introduction

Evolution has frequently been seen as a result of the continuous or discontinuous accumulation of small mutations. Over the many years since Darwin, it has been found that simple point mutations are not the only mechanism driving genomic change, for example, plasmids, transposons, bacteriophages, insertion sequences, deletion and duplication, and stress-sensitive mutation all have a part to play [Sha97, Sha99] in directing the genetic composition and variation of organisms [Koc93] towards meeting the moving target that is the environmental ideal at any one time. Considering the probability of single point mutations arising and repair mechanisms that may act to counteract their accumulation, it is unlikely that simple mutation can create rapid diversity. It is clear that evolutionary change depends more on larger scale changes in genomic sequences caused by sexual and other forms of horizontal gene transfer. These generate the variation necessary to allow rapid evolutionary response to changing environmental conditions.

Predictive models of *E.coli* cellular processes already exist, the E-Cell project [THTSMMSTYVH99] aims to use gene data directly in a mathematical model of transcription. The Virtual Cell [SFSCL97, SL99] project makes use of user-defined protein reactions to simulate compartments at the nucleus and cellular level. Gepasi3 [Men97] also models protein reactions, but from within an enclosed box environment. The BacSim [KBW98] project simulates individual cell growth at the population scale. Eos [BSS00] is also based at the population scale, but is intended as a framework for testing idealised ecologies, represented by evolutionary algorithms. These tools and those that they rely on are excellent models of behaviour. However, they suffer the same drawbacks; all rely on actual experimental data to be input and more importantly, once input that data is static. The aim of this study is to answer some of the questions regarding bacterial evolution and the role played by genetic events other than simple point mutation using an evolving multicellular and multispecies model that builds up from the scale of the genome. In effect, it is not bacterial evolution that is being interrogated, but the co-evolution of bacteria and any organism that has a direct effect on the genetics of those bacteria.

To test these questions, it is necessary to build a model that attempts to encompass what are considered the important qualities of bacterial evolution and bacterial life, but is not overly specified as to constrain the results. The model is therefore a careful balance of biological and computational realities [Way01] with an emphasis on open-endedness [Kam96]. The biological literature has many examples of the possible forms of mechanism within the relatively 'simple' example of *E.coli*, but even this must be carefully constrained. It is clear that computational models lack computational power when compared to real world processes.

In focusing attention on aspects of the *E.coli* system, it is clear that there are two new insights provided by the emerging disciplines of genomics and proteomics. Proteomics is the study of enzyme and protein interactions. Traditionally this meant differential equation models of interaction. However, nowadays there seems also to be an implicit link with the application of protein descriptors derived from sequence information in identified genes [KSBCHHHS97], an application that has only recently become tractable with the arrival of accurate genome data. Genomics is the study of genome structure, interaction and encoding and has been stimulated by the Human Genome project [KH01] as well as whole genome sequencing projects for many other organisms, notably those for numerous bacteria. The genome should perhaps regarded not as a book that is continually read from, but rather a program that is continuously executed and adapted over the life time of individual cells, tissues or entire organisms. From this it appears that interactions within cells involve the combined effects of enzymes, structural and regulatory proteins acting on genes, which in turn act on those enzymes and other proteins, creating a huge number of both positive and negative feedback loops necessary for controlled execution [Fre00]. The ideal model therefore is one that takes both these stages into account and allows for the evolution of the genome in the presence of other genomes, each genome being an implementation of what many conceive as the computational cell [AR94, Bra90, Sha91, Bra95, Pat98, DHB00, RLM96].

Looking at this another way, the main aim of COMIC is evolutionary modelling based on biologically realistic organisms. Evolution requires something akin to a genome and the above argument shows a genome by itself is inert and only part of the overall mechanism. There are then three themes to the COMIC model: the environment, the genome and functional proteins, including enzymes.

2 Biology

2.1 DNA, RNA and proteins

As it is widely disseminated, DNA is the carrier of genetic information and stores all instructions for the development of every known living thing. In its simplest representation DNA is simply a chain of the bases adenine, cytosine, guanine and thymine, linked together by phosphodiester (sugar-phosphate) bonds to form a regularly spaced chain. For each DNA strand there is normally an anti-parallel complementary strand that is annealed to it by hydrogen bonding. Cytosine complements guanine and adenine complements thymine (the relationship is reciprocal). As only complementary bases hybridize but bases on the same strand are chained together; the complementary bases point to each other and the phosphodiester bonds run along the outer edges of the strand pair. The complementary strands wrap around each other in a coil, and this coil is itself regularly coiled. To put some numbers to the dimensions of the strand(s), this coil formed by the two coiled strands is 2 nm in diameter and the distance between adjacent bases is 0.5 nm. What is difficult to imagine is the scale of the object, in humans this double helix is then coiled around itself to a diameter of 11 nm, coiled around itself again to a diameter of 30 nm, coiled again to a diameter of 300 nm and then wound in a 700 nm diameter spring like formation. The end result forms the 'sticks' that make up the chromosome *i.e.* 4 sticks joined at the same point would make up an X chromosome, 1400 nm wide and containing 4x106 bases. It is called X because at this scale it is big enough to see under a microscope and really does look like an X; chromato meaning coloured and hence visible.

The *E.coli* genome is around 4.6 Mbp (million base pairs) long and arranged in a circle, composed of many loops each 50 to 100 kbp long joined by proteins. The DNA is supercoiled, and is further compacted by proteins *e.g.* HU, a dimeric protein that binds DNA non-specifically and wraps the DNA around itself.

2.2 Transcription

Part of the central dogma proposed by F.H.C.Crick in 1958 is that RNA is created from DNA by a transcription phase where information from DNA is transcribed into RNA. RNA differs from DNA in that it contains uracil in place of thymine and is frequently single-stranded, including looping back on to itself caused by sections of intrastrand homology. Some reports (*e.g.* [Gil87]) have suggested that RNA could be the initial bases for life. For the purposes of this work however, RNA will be referred to as a message carrier that comes in three basic varieties, messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA rRNA. The transcription process is carried out by RNA polymerase (an enzyme specific for the creation of RNA species on a DNA template). For the purposes of the COMIC simulation, all references to RNA that follow should be regarded as to mRNA.

The first step in the creation of a protein from the DNA is transcription by a DNA-dependent RNA polymerase (a complex protein based machine). RNA polymerase catalyses transcription in a process that requires double-stranded DNA as well as the nucleotides ATP, GTP, CTP and UTP (*i.e.* fuel). Transcription is directional, with mRNA chain growth progressing in the 5' to 3' direction using a DNA template strand with the opposite polarity (3'$\rightarrow$5') (the sense strand). The complementary strand (running 5'$\rightarrow$3') is called the anti-sense strand. In *E.coli* the RNA polymerase moves at 40 bases per second at 37^{o}C, transcribing as it moves.

To start transcription, an RNA polymerase molecule binds to the double stranded DNA, ideally at a promoter site. The RNA polymerase and all required cofactors constitute a transcription complex. The initial base (start) of the transcribed region after the promoter is known as the +1 position and the promoter and any operator sites have negative nucleotide designations negative relative to this position. As with replication, the DNA strands must be unwound to allow transcription both to be initiated and to proceed. Transcription stops at a terminator sequence which contains a self-complementary region that can form a stem-loop or a hairpin structure out of the RNA product. This structure ensures that the transcription complex stops and promotes dissociation of its constituent parts.

2.3 Protein Structure

There are two classes of proteins, globular and fibrous. Globular proteins can be regarded as spherical particles as they are folded compactly. Most enzymes are globular. The fibrous proteins have a high axial ratio (length/width) and are typically used in a structural role. A protein takes a shape dictated partly by the primary polypeptide sequence and is stabilised by a variety of forces that hold it together; hydrophilic side chains tend to the outside and hydrophobic amino acids remain on the inside.

Except for a few catalytic RNA molecules, an enzyme is almost certainly a protein. An enzyme is a catalyst for reactions that would occur but very slowly. COMIC considers them essential, the large difference in reaction rates makes it not worth considering the case when the reaction occurs without the enzyme - as a result COMIC ignores the passive components and just models the enzymes.

This is also supported by another limitation, a complete system would involve the interplay of basic chemical building blocks and physical processes, requiring the implementation of all the chemical and physical processes that can possibly occur. Unlike more fixed simulations, COSMIC must be open ended and so must support a complete range of possible chemical and physical interactions without favouring one set of fine grained interactions over another set. But, as there are more processes than time to implementation them and too many important processes have no predictable function in a evolving environment, COSMIC is deliberately limited to avoid fine grained chemistry and undecidable processes; instead the focus is on the concepts of optional transcription and evolutionary mechanisms.

The encoding from DNA to protein function is ill defined (hence the protein folding problem [Hau97, KSBCHHHS97]. The bases are read three at a time as codons and converted to 20 different types of amino acid that are linked by peptide bonds to make a polypeptide chain. There are 64 potential codon combinations, many more than required to encode all possible amino acids, but allowing inclusion of two start codons, three stop codons and the use of varying degrees of redundancy in codon specificity for each amino acid. Although the base alphabet is 20 letters in size, the polypeptide can form words of several hundred amino acids in length. This combined with the fact that shape (influenced by the encoding) dictates the role of the protein then makes nearly intractable the task of determining the role of a real protein. The shape can take any imaginable 3D form (obviously the chain cannot go through itself), using regular patterns that change over its length.

2.4 Optional Transcription

To further the level of complexity, there is variation in gene expression and control, as enzymes and regulatory proteins may be used to control transcription of a region of DNA. Controlling proteins that are either transcription repressors or activators that may themselves be controlled by other activators.

The process is partly explained by the Jacob-Monod theory, which is generally applied to prokaryotic gene expression rather than eukaryotic gene expression. The theory uses the unit of the operon that is made up of an adjacent group of structural genes known as cistrons (cistron being effectively equivalent to gene in modern usage), preceded by an operator region. The operator site forms a lock on to which a repressor protein (DNA-binding protein) can attach. Once attached and so the proteins (or the various mRNAs) in that region are not manufactured. If an inducer molecule is present, it can bind to the repressor molecule and nullify the effect of the repressor. There is also the case of the corepressor in which the repressor will only bind to the operator region when it has already combined with another repressor of the required type. The regions on the DNA strand can be shown with Z representing the cistrons, P representing the promoter and O representing the operator. Repressing or activating proteins reflect both the genotypic and environmental or genotypic state of a bacterial cell, creating fairly rapid complex processing without a nervous system.

2.5 *lac* operon

The principal idea of optional transcription comes from research on the *lac* operon found in *E.coli*. This is only one heavily researched example using the common idea of optional transcription.

The *lac* operon is concerned with the use of lactose as a carbon source, the enzymes that can use lactose are only created when lactose is available. The *lac* operon consists of the structural genes *LacZ* encoding b-galactosidase, *lacY* encoding a galactoside permease and *lacA* encoding a thiogalactoside transacetylase. b-galactosidase is an enzyme that hydrolyses lactose into galactose and glucose. Galactoside permease aids in lactose transport through the cell wall of the bacterium.

These three genes are encoded side by side in a single transcriptional unit called *lacZYA*. Relative to this there is an operator site O_{lac} between -5 and +21, just after the promoter site P_{lac}. If the operator site binds a *lac* repressor protein then transcription is strongly repressed. The *lac* repressor protein itself is encoded slightly upstream of the *lac* operons promoter in the *lacI* gene, this gene is also part of the *lac* operon.

The *lacI* gene encodes the repressor protein but the protein itself is active only as a tetramer, the gene product only works when in groups of 4. Once this has taken place the repressor has a very strong affinity for the *lac* operator and also a high affinity with non-operator DNA.

The *lac* operator site is palindronic, consisting of 28 bp which read the same starting at either the 5' or the 3' ends, the *lac* repressor has the same symmetry when grouped in a four unit tetramer.

In the absence of lactose, the repressor protein binds to the operator site, though it is thought that this doesn't stop the RNA polymerase from binding and instead just stops its progress. Strangely, the binding of the *lac* repressor to the operator site increases by two orders of magnitude the affinity of the RNA polymerase to the promoter site, making it quite likely that an inhibited operon also has bound RNA polymerase.

When repressed the *lac* operon generates a very low level of gene product. When lactose is present, the low level of expression allows slow uptake of lactose, some of which is converted to allolactose. Allolactose binds the *lac* repressor, changing its affinity for the operator site and so forcing the unbinding of the repressor. As the RNA polymerase will probably be already present, transcription can start immediately. The removal of the lactose inducer leads to a rapid inhibition of transcription, as rebinding of the repressor is almost immediate and the *lacZYA* RNA transcript is very unstable.

The promoter site P_{lac} and other related promoters have a strong affinity for RNA polymerase, the -35 sequence can be weak and even the -10 sequence can be weak. For high rates of transcription initiation, to increase the rate requires a specific activator protein called a cAMP receptor protein (CRP or Catabolite Activator Protein, CAP). CRP exists as a dimer that cannot by itself have any effect on transcription rate. When glucose is absent, the level of cAMP increases and CRP binds to cAMP producing a CRP-cAMP complex that binds to the promoter site slightly upstream from where the RNA polymerase would bind. The DNA is bent by the presence of CRP, forming a 90^o bend which is believed to multiply RNA polymerase binding affinity by 50-fold. In practice, the location of the CRP binding site can vary much more between operons than stated here, the site can be on the promoter, next to the promoter or be much further upstream.

2.6 *trp* operon

The tryptophan operon encodes five structural genes which are required for tryptophan synthesis. The RNA transcript produced is a single 7kb long strand, starting from the operator site O_{trp}. As with the expression of the *lac* operon, the RNA product is unstable and so regulation of DNA quickly regulates the protein end product, which in this case is tryptophan.

The *trpR* operon is the source of the *trp* repressor and is located somewhere upstream of the *trp* operon. The operator sequence is symmetrical and forms the repressor binding site, which also overlaps with the *trp* promoter site between bases -21 and +3. The core repressor binding site is a palindrome 18 bp long. The *trp* repressor only actively binds the operator site when it has itself has formed a complex with tryptophan. The repressor is a dimer and has a structural similarity to CRP protein and the *lac* repressor, the dimer needs two tryptophans to be complete. It is the tryptophan that gives the dimers structure the correct distance between its two reading heads and its central core.

The five structural genes encode for enzymes that produce tryptophan, tryptophan therefor inhibits its own synthesis by a magnitude of 70 (presumably under artificial tryptophan simulating conditions), which is much smaller than that caused by binding of the *lac* repressor.

So far the *trp* operon seems like the *lac* operon, but for the self inhibition. As well as the normal transcription controls there is also an attenuator sequence following a leader sequence using around 162 nt before the first structural gene *trpE*. This attenuator is a short area rich in palindromic GC bases followed by each U bases and it is also rho-independent. If this sequence manages to form a hairpin structure in the transcribing RNA, then it will act as a terminator and force early termination at around 140 bp long, stopping before the structural genes have been reached.

The leader itself also has a role to play; divided into 4 successive sequences, 1 and 2, 2 and 3, and 3 and 4 are complementary, and so can bind to themselves to form a hairpin which stops further RNA transcription. If 2 and 3 bind then this does form a hairpin but does not stop transcription. Under normal conditions the binding of 1 to 2 and 3 are 4 are more favourable than 2 to 3.

Also in this leader is an efficient ribosome binding site and successive codons encoding for tryptophan. Under conditions of low tryptophan availability the ribosome would pause at this point. Since transcription and translation are tightly coupled in *E.coli*, the net effect then is to negatively control tryptophan transcription. In reality the hairpin formation between sequences 3 and 4 is more likely when tryptophan level is high, the pause occludes sequence 1 leaving sequence 2 to bind with sequence 3. In the alternative case, the pause occurs at the start of sequence 2 and so sequence 2 is occluded allowing sequences 3 and 4 to form a hair pin.

Given both forms of optional transcription, the tryptophan dependent repressor and the tryptophan dependent attenuation region, total tryptophan level can be amplified by 700 times. The attenuation sequence giving a 10 fold increase and the tryptophan dependent repressor giving a 70 fold increase. Generally, attenuation is present in at least six other operons with a role in amino acid synthesis. For example the *His* operon, but in this case the attenuation mechanism is the only means of control, there is no operator.

2.7 An *E.coli* environment

It is clear that prokaryotes, including *E. coli* rely on horizontal transfer of genetic information [OLG00], this comes about both from contact with DNA, by cell-to-cell contact (conjugation) and through infection by bacterial viruses (bacteriophages) occasionally carrying non-viral DNA. An environment is clearly needed to simulate bacteria, phages and 'environmental DNA'. An environment is also a useful tool with which fitness can be implicitly defined. Fitness is defined here as a measure of genome convergence towards the ideal. However, the changing nature of the environment ensures that there is no single ideal and that any better solutions may become inferior over time.

More immediate genome fitness is obtained by direct response to the environment, the environment is enriched with glucose in a patchy distribution and each cell has a number of receptors that transport glucose to the cell interior. The genome in combination with the proteome normally responds with activation of flagella biosynthesis allowing motility by a swimming action [Mac96]. COMIC doesn't synthesis the flagella, simulation time can be shortened by assuming they exist; instead it is glucose that activates the rotation of flagella.

3 The Genome and the Proteome

On the surface, a model genome would seem to be best represented by a simple string of letters. The literature shows that although this is true, there are layers of interpretation that can be placed on top of the simple representation. Looking at genome maps (http://www.ecocyc.org) shows that the strings are divided into non-uniform lengths, each of these identifying some gene or other active string sequence. COMIC considers the genome to follow this representation with the DNA string representation just under the surface as a method of assigning gene labels and mutually reactive enzymes. Sequences can be broadly categorised into those that encode a protein and sequences that act as regulatory structures on which proteins (or further nucleic acid sequences) act [RRCMS96, CGB98]; the reality is that these types overlap, and multiple subtypes can be formed from the same sequence. As a result of this, the genome in COMIC dynamically encodes 8 broad sequence types and allows for mixed types of single genes. Sequence type then follows from pairing fixed types such as the interface to the environment and control sequences, to dynamic types defined by strings that were originally defined to be only genes. Sequence pairing is based on the anti-match of the DNA like encoded strings and so responds to the changing genome but is also consistent over time.

Typically, models of differential equations provide the practical simulation metaphor for biological simulations. At present, it is only E-Cell that incorporates both the genome and the proteome. Differential equations are however very much akin to statistics: they represent mean quantities and ignore the irregularities of the system in question. COMIC uses the individual-based modelling approach in which each molecule is considered as an individual entity with individual parameters. This allows for the inclusion of spatial effects [RGW01] without imposing artificial spatial boundaries of fixed resolution. Given a population of functional polypeptides in the cell, each has a chance of reacting with each other and on the genome. The probability involved is based on their type as defined by the genome, their position, their half-life and their age. Each reaction lasts for a variable time based on the type of pairing. At this point, it must be pointed out that simplification of the real biological process has lead to COMIC having no explicit molecule type on which enzymes react. A single type of generic protein is assumed to exist and all reactions act on that pseudo protein. This protein can be bound by enzymes, as well as the normal process of enzymes acting on each other; the resulting effect is that the enzymes and the single metabolite type is not available to do anything else while bound, allowing another molecule a chance to bind. Another simplification is in the area of motility, there is a direct relationship between motile response and optional transcription. Chemical simulation is beyond the scope of COMIC and yet flagella motion is primarily created by proton motive forces over the cell wall [Mac96], coming from both electrical potential and chemical potential. As a result, this simplification seems justified. As an aside, the developmental genetics of flagella are well described, but this aspect is also beyond COMIC.

It is hoped these necessary simplifications have little impact on the evolutionary results obtained with COMIC, the primary point of enquiry is the change in the genome over the course of medium and long term time; the genome and the concept of the operon [RRCMS96, TMBW97, Wag00] is intended to give the COMIC population their adaptability to respond and adapt; the completeness of genome operators was therefore seen as important, gene types such as operators, promoters, repressors, anti-repressors, terminators, RNA polymerase and metabolites form the main classes. These types should give the genome the ability to express any function required without the need of realistic metabolites.

4 Model

The fully specified COMIC model consists of a hierarchy of sets or objects. The most basic set is the gene, which is part of a genome set, which is part of a cell set as part of the environment. Each level contains additional parameters associated with that level, for instance genes and genomes contain spatial information that partly compute reaction probabilities.

From the genes come gene products, a single gene can lead to many gene products so with each gene is potentially a set of gene products. Each gene product contains spatial parameters and a calculated time to live (the term half-life seems inappropriate when describing an individual rather than a population).

Gene products can react with other gene products and operate directly on the genome, decrypting this relationship requires several steps. Firsty, because the relationship between gene and protein function are currently intractable (especially for completely unknown genes), function mapping has been constrained to use only a direct hamming distance between genes. A gene (which always leads directly to its gene product) can react with another gene if the hamming distance is below a specified threshold. When specified for all genes, these relations give a picture of which gene product (enzymes) can react. Not all sequences are genes, control sequences are also considered in the hamming process, this creates interaction paths between control sequences and gene products - thus allowing these gene products to then be called sigma factors, repressors and inducers. A summary of possible interactions, sources of transcription products and a map of indirect or direct attenuation is shown in figure 1.

Given a genome comprised of genes, the gene products of each and a set of

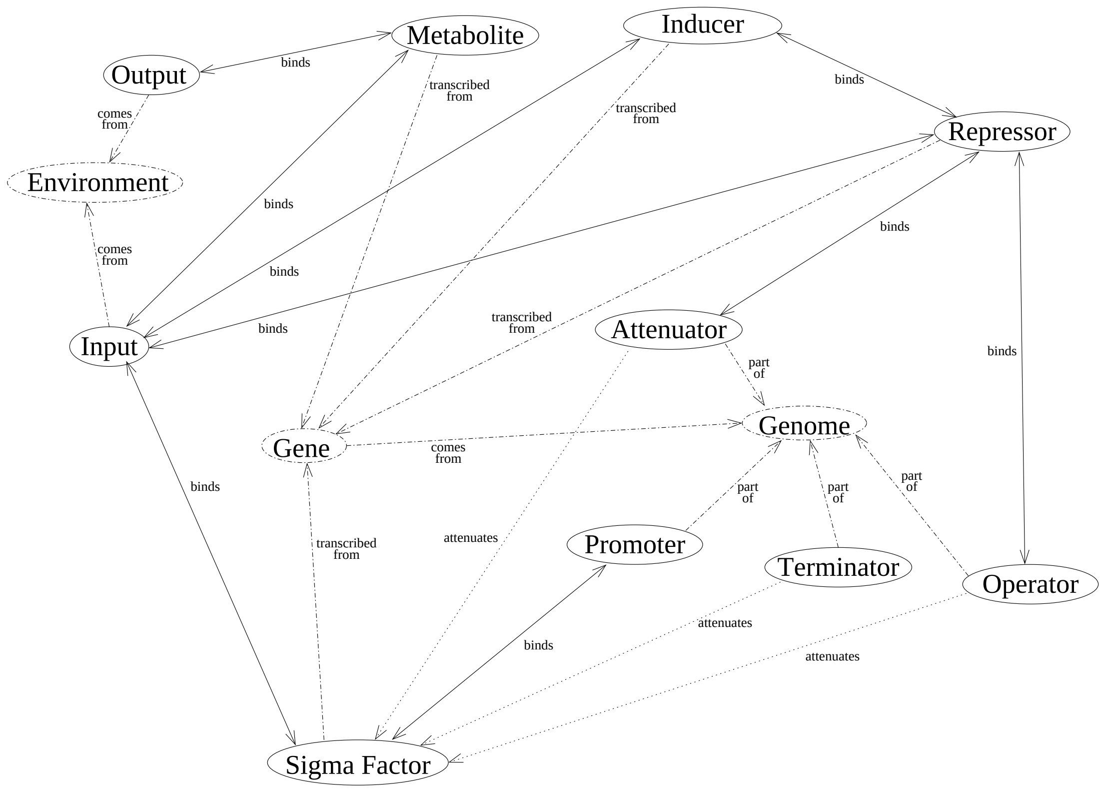

Figure 1: Enzyme Interactions

relations describing what can happen, the only missing set is of reactions that are happening. This is based on the relations of possible reactions, each relation carries with it a set of reaction instances. Each instance carries with it references to the specific enzyme instances involved as well as a mutual position in space and a reaction duration based on the type of relation.

The only aspect missing from this system are sense/response pathways. For this COMIC considers receptors on the cell wall (input) and flagella motion (output) to be directly related to the transcription pathways. The receptors represent pseudo gene products that exist in a fixed position. These can bind to normal gene products using a parallel set of relations also based on hamming distance.

The flagella response (output) follows in the same way. In both cases, the effect is not one of convertion, *i.e.* a receptor places a protein in the cell and that goes on to be converted to something else by gene products - crossing the protein and chemical scale was never the remit of COMIC. Instead, the mode of operation follows that of use, if a gene product is bound at one location it can not also be involved in a reaction elsewhere. This is quite a simplification but as the role of COMIC is mainly in evolution it seems justified.

The largest scale of COMIC is the environment. To test evolution, a population of cells must compete side by side for what exists in a common environment. To this end all cells live in a glucose rich environment that is depleted over time by the cells. In common with Genetic Algorithms, better cells in better regions of

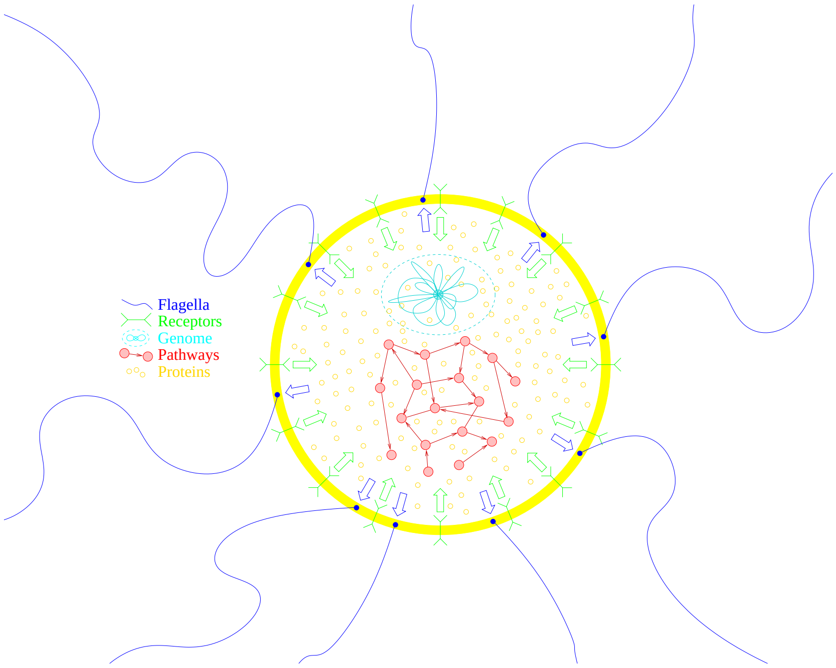

Figure 2: Conceptual drawing including all high level model attributes but without the detail of the underlying parameters.

the environment will grow faster and so multiply faster. For this COMIC uses quantitatively accurate cell growth equations [KBW98, NTT96, BD96].

The combination of input reward based on cell position and cell position based on flagellum output produces an indirect reward based system that is the basis on which the simulated *E. coli* evolve.

Evolution and other causes of cell death are then supplied by the other organisms and mutations in the form of both simple point mutation and the longer gene sequence duplication and gene sequence deletion. This creates an ecology where it is good to be fit and also good to be genetically unique [KD87, DPA96], an ecology such as this also allows for a dynamic population size.

5 Results

The COMIC system clearly has two distant scales, the environment state and the state of each cell. COMIC can generate a vast amount of detailed raw data on the environment and more importantly, the gene expression of each cell coupled with gene linkage maps. The gene expression data shows that even random loosely connected genomes can achieve a wide range of transcription rates and connection to the environment.

COMIC can be considered a batch process but for the never-ending nature of the batch. The architecture ultimately allows a never-ending simulation in which state can be recorded and reloaded why changing the global and local parameters.

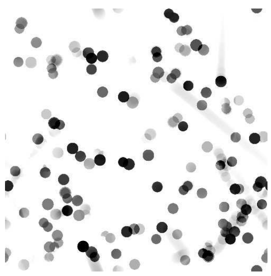

Figure 3: Early Environment

The current COMIC produces a variety of common visualisations, the common feature is the abstract nature of the labels in that real proteins are not simulated and so data can be isomorphic but not easily comprehensible. These being per cell gene expression charts and pathway di-graphs. Also generated are per cell graphed averages of major parameters and snapshots of the environment at the population level. For unconverged and therefore random genomes such as these, most however make little sense without careful individual analysis.

5.1 Environmental Macroscopic View

Figure 3 shows environmental change after 4000 seconds of simulation. White equates to a glucose level of 4.5mg, black equates to none. This area is 0.2mm square. Black circles represent bacteria which haven't moved through lack of connection with their flagella, grey streaks show moving bacteria - bacteria cannot move without leaving some visible trail. Per cell glucose use has been exaggerated to better motivate evolutionary change.

The genomes are random with only a small degree of mutation and so intelligent behaviour can not be observed, what can be seen is cell 177 moving at high speed around the environment, cell 42 moving slowly and the majority of cells not moving at all.

5.2 Cell Lineage

Lineage charts collect together related cells in a hierarchy, starting with the original parent cell and its random genetics through related generations which inherit the parent's entire genome and on average 50% of the enzymes.

Figure 4 shows cell 42 lineage, initially seeming to struggle for life but then dividing to create cell 177 and then something obviously changes, as 42 dies but 177 goes on to produce a large population of related cells. This lineage still exists several hours later when the simulation was stopped, having 108 cells of the same lineage. The annotation shows the cell number when spontaneously created (above the lineage line), the cell number of the death (below the lineage line), the division of cells (to the right of the graph) and the left hand side showing absolute time (simulated seconds) with time difference to previous event. Also, to help follow

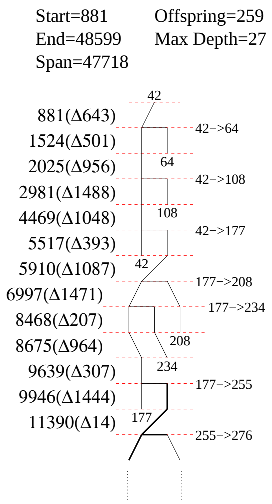

Figure 4: Cell Lineage for cell 42

the same line, every fifth cell line is bolder than the others.

5.3 Gene Expression

COMIC produces gene expression charts for all possible expressible genes at a maximum time resolution of Δt seconds, however a more useful resolution of 1 is shown here. This picture shows gene expression for 82.5 minutes from initialisation, 15 minutes per row (separated by white borders). Colours and shades of colour indicate the levels on a logarithmic scale, shades of blue being the lowest, green being the highest and gray being unused space that compensates for the change in genome size over time. Analysis of this expression data is currently limited to manual inspection, indicators being gene usage, the change in expression and the more obvious cases of gene insertion/deletion.

Figure 5 shows Cell 42, the initial cell of the long running lineage. As can be seen from the first row, there were many gene duplication events, leading to an overall enlarging of the genome. Gene expression took some time to settle down and never fully converged on any one pattern. This could only be a good sign that the relatively small changes brought about through mutation are having a large effect on the proteome. There are 3 cell divisions hidden here, at 10.7, 35 and 72.3 minutes, the latter being cell 177.

Figure 6 shows Cell 177, the third child of cell 42 and what later become the parent of a large number of cells. As can be seen this cell is also constantly changing, though the speed of change is not so dramatic. Comparing the start and the end shows how much the new cell changed from its parent, it can also be seen how much larger the genome grew despite growth and reduction having the same probabilities.

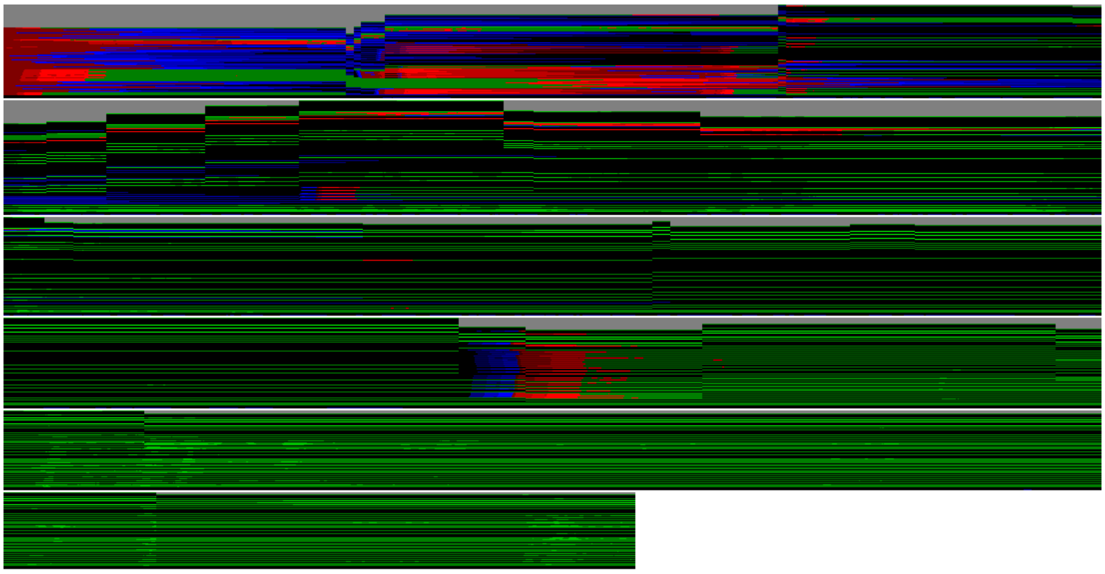

Figure 5: Cell 42 gene expression

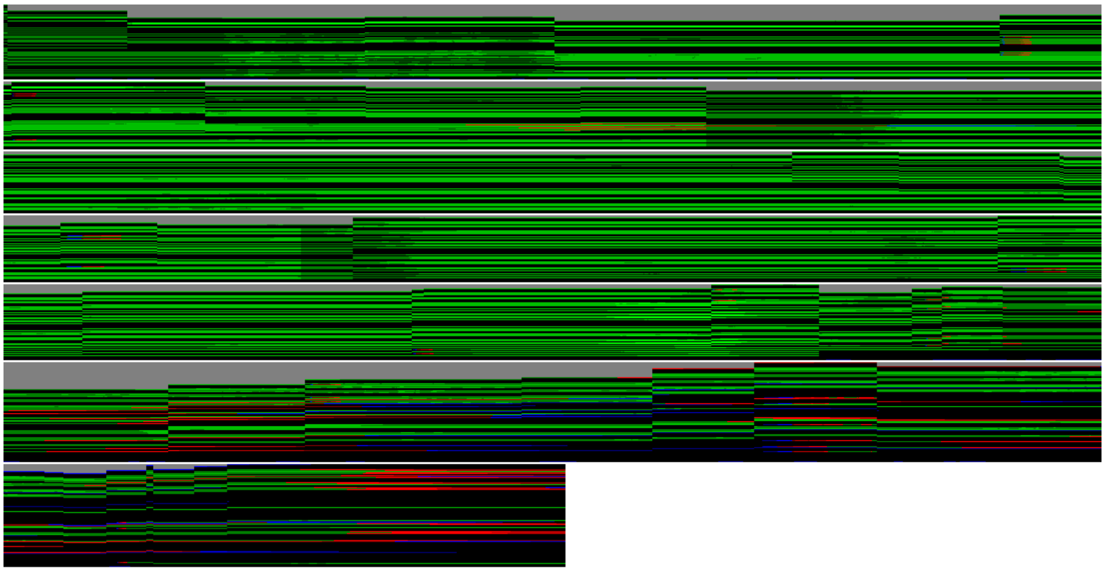

Figure 6: Cell 177 gene expression

5.4 Pathway Digraphs

COMIC also generates digraphs representing interaction pathways within each cell. Shown in figure 7 is a fragment of cell 42, the genome is random and so there is no real structure. Nodes show genes and pseudo genes of input/output receptors. Edges show relationships between these interactions, either transcriptional relations (adjacent genes) or control relationships (all the types shown above). Inside each node is the gene code and the current type associated with this gene. Figures on the edges show total usage counts for binding and unbinding reactions.

5.5 Cell statistics

5.5.1 Statistics for cell 42

Shown in figure 8 are some overall statistics for cell 42. The top graph shows cell energy, used to constrain futile cycles. As can be seen, cell energy is slowly depleting. The next graph shows cell mass, starting from some value all cells increase to 0.4 fl then divides, growth being dependent on substrate. Three cell divisions can clearly be seen. The next two graphs show the cell x/y position in the environment (as shown in environment view figure). Both graphs show the cell is moving at some speed and so must be covering fresh substrate regularly - an initial requirement of a converged cell. The third graph shows total enzyme population over time. As the enzyme population is divided when the cell divides, the approximate halving of cell contents can be seen coinciding with halving of cell mass. The penultimate graph shows receptor activity over time, that is the ability of the cell's optional transcription network to bind with glucose receptors on the cell wall - it is regarded as the cell input. As can be seen, there is some activity but this would never be enough to guide the cell intelligently around the environment. The final graph shows flagella response, the equivalent of cell output. This shows the ability of the pathway network to interact with the flagella and so move the cell. As can be seen (largely by the x/y graphs) there is sufficient cell movement to achieve near maximal growth, the important point here is that output lacks control, this unconverged cell keeps the cell moving regardless of input.

5.5.2 Statistics for cell 177

Graphs in figure 9 show the same overall parameters as previously but for cell 177, the third child of cell 42. The most obvious difference between the two is the number of enzymes, in this case it is now relatively constant; quickly recovering from cell divisions. The unusual case that appears in so many COMIC cells is at time $9.5x10^3$, there is a rise in the enzyme population, a fall, a halving caused by cell division but then the cell never really recovers and looses a connection with the environment (both input and output), COMIC then eventually kills the cell as it is considered no longer viable.

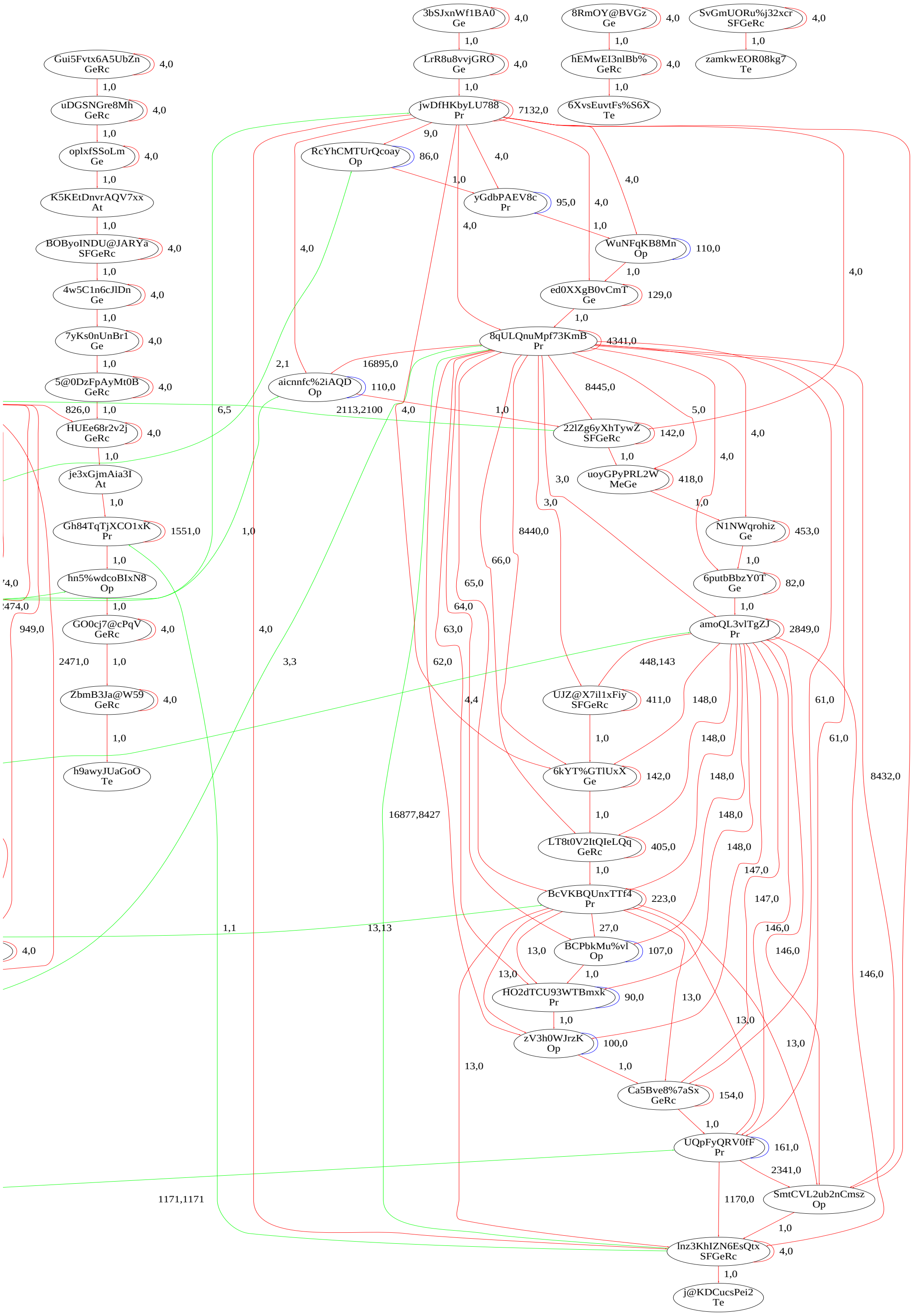

Figure 7: Interaction Pathways

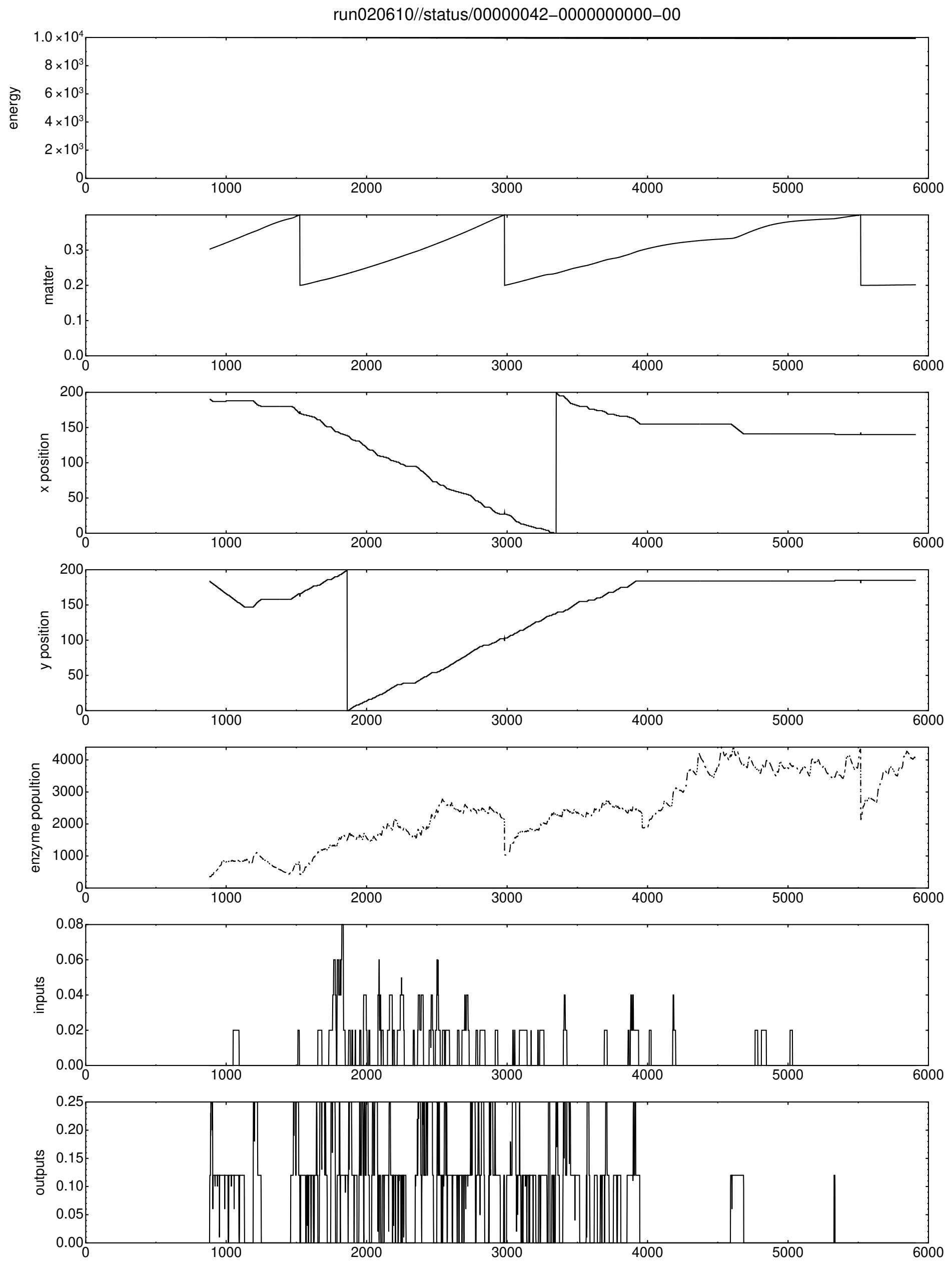

Figure 8: Cell 42 statistics

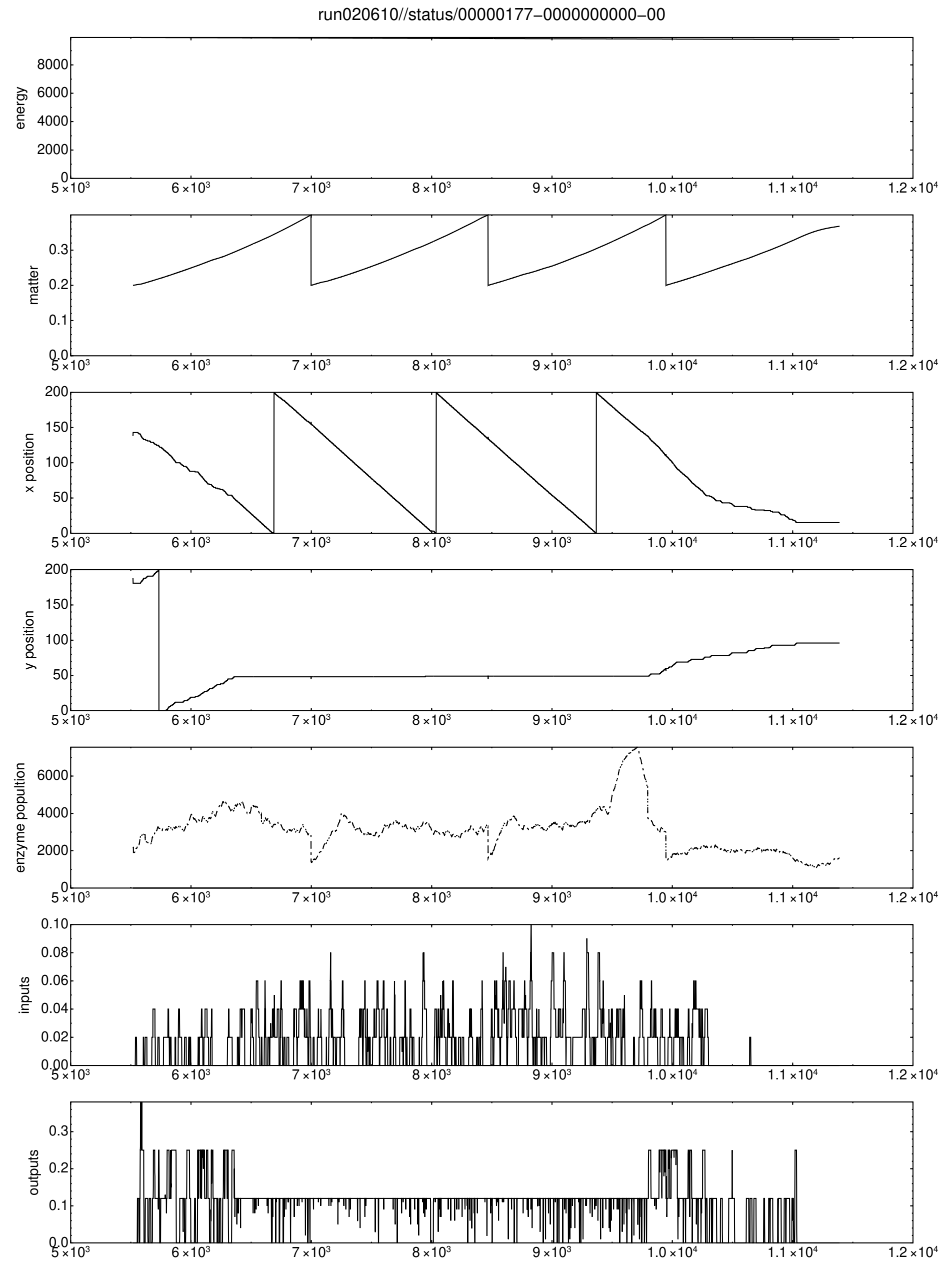

Figure 9: Cell 177 statistics

6 Discussion

The COMIC model is a growing tool with which evolution can be modelled in a greater detail than ever before allowed. The problems this brings are twofold, the computational effort is significant and stands to limit the evolution of COMIC itself, however the biggest hurdle is the shear amount of data generated. The individual based philosophy is clearly then a double-edged sword, escaping global averages means always considering individuals. This leads on to further stages of COMIC development, tools to analyse the output in a meaningful and concise way.

References

[AR94] Adam Arkin, John Ross (1994). Computational Functions in Biological Reaction Networks. In *Biophysical Journal.* Volume 67, pp. 560-578.

[BD96] Hans Bremer and Patrick P. Dennis (1996). Modulation of Chemical Composition and Other Parameters of the Cell by Growth Rate. In *Escherichia coli and Salmonella: Cellular and Molecular Biology*, F.C. Neidhardt *et al* (Eds) Volume 2. 2nd Ed. ASM Press. pp. 1553-1569.

[Bra90] D. Bray (1990). Intercellular signalling as a Parallel Distributed Process, *Journal of Theoretical Biology,* 143, 215-231.

[Bra95] D. Bray (1995). Protein molecules as computational elements in living cells. *Nature, Vol. 376. July.*

[BSS00] E. Bonsma, M. Shackleton and R. Shipman (2000). Eos - an evolutionary and Ecosystem research platform. In *BT Technology Journal* Vol 18, No. 4. October 2000, pp. 24-31.

[CGB98] J. Collado-Vides, R.M. Gutierrez-Rios, G. Bel-Enguix (1998). Networks of transcriptional regulation encoded in a grammatical model. In *BioSystems*, V47, No.1,2. pp. 103-118.

[DHB00] Zhenhua Duan, Mike Holcombe and Alex Bell (2000). A logic for biological systems. *BioSystems 55*, pp. 93-105.

[DPA96] Devine, P., Paton, R. & Amos, M. (1997) Adaptation of Evolutionary Agents in Computational Ecologies. In *Biocomputing and emergent computation: Proceedings of BCEC97.* World Scientific, Lundh, D., Olsson, B. & Narayanan, A. (eds.), ISBN 981-02-3262- 4, 66-75.

[Fre00] Matthew Freeman (2000). Feedback control of intercellular signalling in development. In *Nature*, volume 408. pp. 313-319.

[Gil87] W. Gilbert. The Exon Theory of Genes, *Symposia on Quantitative Biology, Vol. 52, 1987.*

[Hau97] David Haussler (1997). A Brief Look at Some Machine Learning Problem in Genetics. In *Proceeding of COLT 97.*

[Kam96] George Kampis (1996). Self-modifying systems: a model for the constructive origin of information. In *BioSystems*, Volume 38, pp. 119-125.

[KBW98] Jan-Ulrich Kreft, Ginger Booth, and Julian W.T. Winpenny (1998). BacSim, a simulator for individual-based modelling of bacterial colony growth. In *Microbiology,* 144, pp. 3275-3287.

[KD87] J.R.Krebs and N.B.Davies 1987. An introduction to Behavioural Ecology. Blackwell Scientific Publications.

[KG98] Marc Kirschner & John Gerhart (1998). Evolvability. In *PNAS*, Volume 95. pp. 8420-8427.

[KH01] W. James Kent, David Haussler (2001). GigAssembler: An Algorithm for the Initial Assembly of the Human Genome. *In press.*

[Koc93] L. Koch (1993). Genetic Response of Microbes to Extreme Challenges, *Journal of Theoretical Biology 160, pp. 1-21.*

[KSBCHHHS97] Kevin Karplus, Kimmen Sjölander, Christian Barret, Melissa Cline, David Haussler, Richard Hughey, Liisa Holm, Chris Sander (1997). Predicting protein Structure using hidden Markov models. In *Proteins: Structure, Function, and Genetics.* Pages 134-139, Supplement 1.

[Mac96] Robert M. MacNab (1996). Flagella and Motility. In *Escherichia coli and Salmonella, cellular and molecular biology.* Second Edition, Volume 1, pp. 123-144.

[Men97] P. Mendes (1997). GEPASI: a software package for modeling the dynamics, steady states and control of biochemical and other systems. In *Trends in Biochemical Science*, Volume 22, pp. 361-363.

[NTT96] Oense M. Neijssel, M. Joost Teixeira De Mattos and David W. Tempest (1996). Growth Yeild and Energy Distribution. In *Escherichia coli and Salmonella: Cellular and Molecular Biology*, F.C. Neidhardt *et al* (Eds) Volume 2. 2nd Ed. ASM Press. pp. 1683-1692.

[OLG00] Howard Ochman, Jeffrey G. Lawrence and Eduardo A. Groisman (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, 18th May 2000, pp. 299-304.

[Pat98] R. Paton (1998). The Ecologies of Hereditary Information. In *CYBERNETICS & HUMAN KNOWING A Journal of Second Order Cybernetics*, Autopoiesis & Cyber- Semiotics 5, 31-44.

[RGW01] Debabrata RayChaudhuri, G. Scott Gordon, & Andrew Wright (2001). Protein acrobatics and bacterial cell polarity. In *PNAS*, February 13, 2001. Volume 98, Issue 4. pp. 1322-1334.

[RLM96] Venkatramana N. Reddy, Michael N. Liebman, Michael L. Mavrovouniotis (1996). Qualitative analysis of biochemical reaction systems. In *Computers in Biology and Medicine 26 (1)* pp. 9-24.

[SFSCL97] James Schaff, Charles C. Fink, Boris Slepchenko, John H. Carson, Leslie M. Loew, (1997). A General Computational Framework for Modeling Cellular Structure and Function (1997). *Biophys. J.* 73, pp. 1135-1146.

[RRCMS96] M. Thomas Record Jr, William S. Reznikoff, Maria L. Craig, Kristi L. McQuade & Paula J. Schlax (1996). *Escherichia coli* RNA Polermerase (E$\sigma^7$0), Promoters, and the Kinetics of the Steps of Transcription Initiation. In *Escherichia coli and Salmonella: Cellular and Molecular Biology*, F.C. Neidhardt *et al* (Eds) Volume 2. 2nd Ed. ASM Press. pp. 792-821.

[Sha91] J.A. Shapiro (1991). Genomes as smart systems. In *Genetica*, Volume 84, pp. 3-4.

[Sha97] J.A. Shapiro. Genome organization, natural genetic engineering and adaptive mutation, *Trends in Genetics* 13 (3) 1997. pp. 98-104.

[Sha99] J.A. Shapiro. Genome System Architecture and Natural Genetic Engineering in Evolution, *Annals of New York Academy of Sciences*, May 1999, 870, pp. 23-35.

[SL99] J. Schaff, L.M. Loew (1999). The Virtual Cell. *Pacific Symposium on Biocomputing*, 4. pp. 228-239. *http://www.nrcam.uchc.edu/publications/vcell_publications.html*

[THTSMMSTYVH99] M. Tomita, K. Hashimoto, K. Takahashi, T. Shimizu, Y. Matsuzaki, F. Miyoshi, K. Saito, S. Tanida, K. Yugi, J.C. Venter, C. Hutchison (1999). E-CELL: Software environment for whole cell simulation. In *Bioinformatics*, Volume 15, Number 1, pp. 72-84.

[TMBW97] P.C. Turner, A.G. McLennan, A.D. Bates & M.R.H. White. *Instant Notes in Molecular Biology.* BIOS Scientific Publishers.

[Wag00] Rolf Wagner (2000). Transcription Regulation in Prokaryotes. *Oxford University Press.*

[Way01] Eileen C. Way (2001). The role of computation in modeling evolution. *BioSystems* Volume 60 (2001) pp. 85-94.