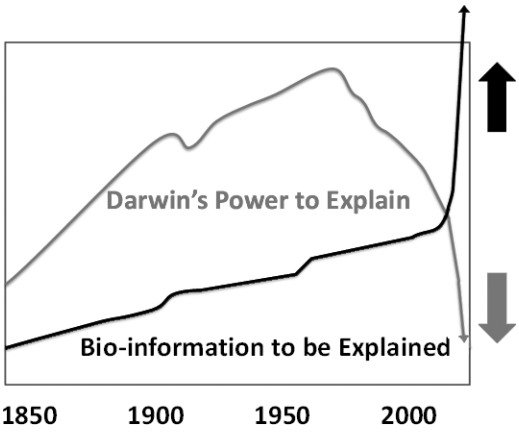
Section Two — Biological Information and Genetic Theory: Introductory Comments

John C. Sanford — Section Chairman

In the 21st century, *biological information* has become the over-arching theme which unifies the life sciences. In the 19th century, Charles Darwin and his colleagues did not yet have the notion of biological information. Indeed Darwin completely misunderstood the nature of inheritance, which he pictured to be Lamarckian in nature. One of Darwin's contemporaries, Gregor Mendel, discovered that the determinants of certain biological traits are transmitted from generation to generation in discrete packages (this work was ignored for a generation). Mendel probably had some vague notion that these genetic packages somehow might contain a very simple type of "biological information". But he could never have guessed that these genetic units which he observed were actually precisely-specified instructions, encoded by language, with each gene being comparable in complexity to a book. When the early population geneticists developed their models, they employed over-simplified mathematical models to try to describe their understanding of genetic change, but at that time genes were considered to be merely "beads on a string."

When DNA was discovered, it finally became clear that genetic information is very much like human written information — an extensive array of language-encoded strings of text. Where did all these text strings come from? For most biologists the already-ruling Darwinian paradigm seemed to be sufficient — they assumed that all biological information must arise merely by random letter changes in the text, combined with some reproductive filtering. In the last 60 years, many thousands of scientists have made a truly monumental effort to try to explain the entire biosphere, just in terms of random mutations which are filtered by natural selection. Has this effort been successful? It has certainly been successful in a sociological sense — this view is now faithfully upheld by the large majority in the academic community. The neo-Darwinian paradigm literally saturates the content of most biological journals. In fact any deviation from this view is generally regarded as academic treason — often being characterized as a threat to science itself. Yet in this section of our proceedings (Biological Information and Genetic Theory), we will show that there are huge genetic problems which bring this reigning paradigm into serious question.

As the figure below graphically illustrates, a paradigm shift appears to be imminent. This is because the amount of biological information which demands explanation is exploding, even while the explanatory power of Darwin's mechanism of



natural selection is virtually collapsing. This section of our symposium focuses on these two things — the explosion and the collapse.

The first problem is the explosion in the amount of biological information which requires explanation. We now realize that the last century's simplistic concept of biological information ("DNA makes RNA makes protein makes life") was incredibly naïve. We are just beginning to understand that biological information is profoundly multidimensional and moves in all directions through elaborate communication networks. The many layers of biological information are not only dynamic, they are globally integrated — overwhelming the previous generation's understanding of information (a gene encodes a protein). This will be clearly demonstrated by **Wells** in the first paper in this section, and is further developed by **Seaman** and **Johnson** in the last two papers of this section. Seaman and Johnson both correctly characterize the cell as being more like a network of computers than a set of books. These papers by Wells, Seaman, and Johnson act as the 'bookends' for this collection of research papers.

We need to better grasp the full scope of what "biological information" really is. It is a serious error to think of biological information as simply the genome. As discussed by Seaman, we can best understand the genome as the hard drive of the cell — it largely reflects *stored static information*. In that light, we should see that the RAM or active memory of the cell is that galaxy of RNAs and proteins which comprises the active communication network within the cell. These RNAs/proteins are actually the *active information* which makes life alive. As discussed by Johnson, RNA and proteins can be viewed as actively operating algorithms, specifying their own folding, their own transport, their own operation, and their various communication links with other molecules. Countless messages are continually being transmitted in both directions between the hard drive (the genome), and the

RAM (RNA and proteins). There is also continuous information being exchanged between different parts of the genome, and between RNAs and proteins, so there is a continuous interchange of information between all components. All this information which is continuously being exchanged within a single cell has been termed the "interactome", and it is vastly more complex than the genome itself. Such interactions within a living cell are beyond counting — and might best be compared to an internet system. The entire cell can be considered to be an extensive communication network. Above and beyond the individual cell, there is still more biological information being regularly communicated between cells, between tissues, and between individuals. Lastly, there is the biological information network that constitutes the brain/mind — which dwarfs everything else we have spoken of. With all this in mind, in this section we will primarily focus our attention on just the simplest level of biological information — the genome.

For decades it was believed that there is just one genetic code, and that only the protein-coding sequences within the genome were functional (less than 2% of the human genome). Essentially all other sequences were designated "junk DNA". This concept has been dramatically reversed in the last ten years, as revealed by Wells in this section's first paper. It is now clear that most of the non-protein-coding genome is functional. This means two things — firstly it means there is a lot more information in the genome that needs to be explained, and secondly it means there are many codes other than the amino acid code.

The implications of having many languages (genetic codes) in the same genome are staggering, and the fact that these codes overlap extensively is breath-taking (see Montanez *et al.*, in the previous section of these proceedings — Biological Information and Information Theory). In addition to the basic protein code, other codes associated with the conventional gene concept include the 12 codes of Trifanov, the transcription codes, the alternative splicing codes, and the RNA folding/processing codes. On an entirely different level, there are genome-wide codes that transcend the gene concept. These include the isochore codes, the nucleosome-positioning codes, the topological 3-D codes, and the epigenetic codes. Even the tiny but super-abundant Alu elements in the human genome, the most famous class of "junk DNA", are now known to contain multiple codes. These include transcription-regulating code, protein-binding code, and also a special 'pyknon' (small RNA) code. Some, but not all, of these codes are described in more detail by Wells. It should be obvious that more codes are waiting to be discovered. In the second to last paper in this section, Seaman, discloses very exciting new evidence for repeat-based codes in the genome, which have an uncanny resemblance to the repeat structures characteristic of executable computer code.

How many genes are in the human genome? The textbooks still suggest there are just over 20,000 human genes — because they have not yet acknowledged the

paradigm shift ushered in by the ENCODE project. We now know that what we used to call a gene was a gross over-simplification. What we used to call a gene, we now know is actually a complex of many functional elements, encoding multiple proteins and many RNAs. If we define each of these functional elements as a gene, there must be hundreds of thousands of genes. Since there is now strong evidence that SINES and LINES are themselves functional elements, we should also recognize these as genes — so depending on how we define a gene, there may be over a million genes in the genome. Our awareness of biological information, just within the genome, is truly exploding. In the following section of this symposium (Biological Information and Molecular Biology), Dent and Wells each present papers proposing additional new types of biological information which entirely transcend the genome. If validated, each of these will clearly require its own language or code. I am convinced that none of us has yet fully absorbed the significance of what is emerging, in terms of the richness and depth of biological information. There has simply never been a more exciting time to be a biologist.

The second problem is the collapse of the Darwinian mechanism, in terms of its power to explain how all this biological information arose and is sustained. This will be clearly demonstrated by the papers of Gibson *et al.*, Sanford *et al.*, Nelson *et al.*, Brewer *et al.*, and Baumgardner *et al.* Natural selection obviously works, the problem is it does not appear to be capable of performing as advertised. These papers show that, most fundamentally, the Darwinian mechanism cannot consistently create a net gain of information. This is because even as rare beneficial mutations arise (only some of which can be selectively amplified), many more deleterious mutations must be accumulating continuously. Certainly this should result in "genetic change over time" — but the change should primarily be downward. If mutation/selection causes genomes to primarily go down, not up, then the Darwinian mechanism cannot explain the origin of genomes, or even their maintenance. Consequently, the explanatory power of the Darwinian mechanism appear to be limited to the trivial and the mundane (i.e., minor superficial adaptations in response to environmental change — mere fine-tuning). This is clearly documented in the following papers.

Gibson *et al.* summarize their extensive numerical simulation research addressing the problem of deleterious mutation accumulation — as affected by the *selection threshold* phenomenon. They have developed what is clearly the most advanced numerical simulation for modeling mutation accumulation within populations ("Mendel's Accountant"). They use numerical simulation to demonstrate that given biologically realistic conditions, natural selection fails to selectively remove the large majority of deleterious mutations. They show that there are various reasons why this happens, but the most important reason in that each population has a certain characteristic selection threshold, and mutations which have

very small fitness effects fall below this threshold, and hence will become essentially invisible to natural selection. Gibson *et al.* show that when biologically realistic conditions are modeled for a higher organism, the selection threshold is especially high, such that the vast majority of deleterious mutations are not selectable, and hence accumulate continuously. If the mutation/selection process is really all that is happening, then this means that all higher organisms should be continuously accumulating deleterious mutations at a high rate, even when there is strong natural selection pressure — which would logically lead to eventual extinction.

Sanford *et al.* have also studied the selection threshold problem, but they examine how it affects the accumulation of beneficial mutations. They use numerical simulation (again, Mendel's Accountant) to demonstrate that there is a very clear selection threshold for beneficial mutations, and that only a very tiny fraction of all beneficial mutations have a large enough effect to be able to respond to selection. They show that the selection threshold problem is even more severe for beneficial mutations than it is for deleterious mutations. Because it is clear that beneficial mutations are very rare anyway, the fact that only a very tiny fraction of them are selectable means that selectable beneficial mutations should be vanishingly rare. When rare beneficial mutations do occur which are above the selection threshold, they respond to selection beautifully and can be rapidly amplified. This reflects the type of response we see when there is strong selection for something like a bacterial mutation for antibiotic resistance. But these types of rare and isolated events can only explain what is known as microevolution (mere adaptation). Clearly, this type of fine-tuning to some specific environmental factor has no bearing on how genomes might be created or sustained. Sanford *et al.* raise the important question — "What mechanism could have established the hundreds of millions of very low-impact nucleotide sites within any higher genome?"

Nelson *et al.* use the well-known Avida simulation program to show that when Avida is run using biologically realistic parameters, the results are remarkably similar to when similar parameters are used in Mendel's Accountant. For example, when a realistic distribution of mutation effects is employed (the Mendel default setting), both programs show no forward evolution at all, but rather a rapid loss of whatever genetic information was initially present. Conversely, when all mutations have very large fitness effects (the Avida default setting), both simulation programs demonstrate explosive forward evolution. Avida, like Mendel's Accountant, when run with biologically reasonable parameters, shows reverse evolution. Nelson *et al*. go on to use Avida to illustrate something that Mendel's Accountant fails to demonstrate — that there is a clearly defined threshold for establishing irreducible complexity via the selective process, given reasonable probabilistic resources.

The profound difficulties with the classic Darwinian mechanism, as described in the preceding papers, have been known within the population genetics community for decades. The standard response to these problems has been either to ignore them, or to invoke ad hoc models to explain away the problems. These ad hoc models have never been critically examined or properly tested. There are two primary models used to explain why genetic change over time might primarily be upward, rather than downward. The first is what can be called the *mutation count mechanism* and the second is the *synergistic epistasis mechanism*.

Brewer *et al.* use numerical simulation to test the mutation count mechanism model. This model suggests that if selection is strongly directed specifically against those individuals with higher mutation counts, deleterious mutation accumulation can be halted. The numerical simulations of Brewer *et al.* show that this mechanism actually can work, but only when mutation effects are relatively uniform, when there is truncation selection, and where there is sexual recombination. However, numerical simulations clearly show the mutation count mechanism becomes ineffective when *any* of the following are true: 1) there is a distribution of mutation effects which is realistically broad; 2) probability selection is operating; 3) a species reproduces clonally. Few if any situations occur in nature where *none* of three conditions are present, hence the mutation count mechanism cannot be operational in any general sense. Therefore, Brewer *et al.* have effectively falsified the mutation count hypothesis.

Baumgardner *et al.* use numerical simulation to test the synergistic epistasis hypothesis. This hypothesis proposes that as mutations accumulate continuously — they will amplify each other's deleterious effect, so that genetic damage does not increase linearly but rather increases exponentially. It is thought that at some critical point, just one or a few additional mutations will create a profoundly deleterious effect ("the straw that broke the camel's back"). In this way selection might be focused more strongly against those individuals who have a higher mutation count (just as with the mutation count mechanism). This hypothesis is highly problematic, is entirely ad hoc, and it is entirely incompatible with all the normal population genetics assumptions. None the less, this hypothetical mechanism is rigorously tested using numerical simulation by Baumgardner *et al.* It is shown that the synergistic epistasis mechanism fails to halt deleterious mutation accumulation, and consistently accelerates mutational degeneration, just as common sense would dictate.

Given all the theoretical evidence that the mutation/selection should yield to a net loss in functional information, it's very reasonable to ask if there are living systems that actually show this might be happening. This is generally difficult to demonstrate experimentally, because most biological systems change very slowly, especially on the level of the whole genome. **Brewer, Smith, and Sanford** have chosen to study RNA viruses, which have short replication cycles and extremely

high mutation rates, and so they can change rapidly in short intervals of time. They examine such viruses to better understand loss of information in real biological systems. They examine pandemic histories which suggest that some human pandemics involving RNA virus may come to an end because of mutation accumulation leading to *natural genetic attenuation* of the virus. They then do a series of numerical simulations that confirm that based upon known RNA viral mutations rates and based upon the biology of pandemics, a significant fitness decline in the virus should be expected during the course of a typical pandemic. These authors then go on to use numerical simulations to examine what factors might accelerate such natural genetic attenuation. They show that use of pharmaceuticals that are known to enhance the viral mutation rate should be highly effective in reducing both the extent and the duration of pandemics. Other practices which should accelerate genetic attenuation would include reducing inoculum levels during disease transmission (stronger bottlenecking), and reducing titer levels in the infected host (lower selection efficiency).

These papers, along with many other lines of evidence (i.e., see Behe's paper in the following section "Biological Information and Molecular Biology"), clearly show that the explanatory power of the classic Darwinian mechanism is suddenly collapsing. This is happening at exactly the same time that we are being overwhelmed with evidence that the actual amount of biological information that requires explanation is vastly deeper and richer than we could have imagined. Surely this is an exciting time to be a biologist!