Cosmos and History: The Journal of Natural and Social Philosophy, vol. 10, no. 1, 2014

SYMBOLIC AND COGNITIVE THEORY IN BIOLOGY

Seán O Nualláin

Abstract**:** In previous work, I have looked in detail at the capacity and the limits of the linguistics model as applied to gene expression. The recent use of a primitive applied linguistic model in Apple's SIRI system allows further analysis. In particular, the failings of this system resemble those of the HGP; the model used also helps point out the shortcomings of the concept of the "gene". This is particularly urgent as we are entering an era of applied biology in the absence of theory, and indeed an era with a near-epidemic of retracted papers.

There are a few workarounds proposed. One is to add to the nascent field of biosemiotics a more explicit concern with syntax. At the time of writing, Apple is being sued for false advertising of its iphone 4s, with the associated claim that apple had solved many of the problems of natural language processing by computer (nlpbc). The system was bought by Apple from a company called SIRI, and in turn was based on the notion, trumpeted by the prior art in a company called Dejima, that nlpbc could be done by keywords alone.

Yet the hype resembles nothing so much as the misrepresentation of the Human Genome Project (HGP) fed to the media in the glory days at the beginning of this millennium, and it says a lot for the status of scientists in society that they have avoided Apple's fate. In this paper, a short review of several current themes in theoretical and applied biology is first proposed. Then the tensions implicit in the notion that the "gene" is simultaneously to be identified as a unit of inheritance and spatially located over spatially well-defined nucleotides is explored and the notion is found to be incoherent. An expanded notion of inheritance is proposed in the context of a focus on inheritance as necessarily involving species, population and organism over time.

While it is premature to talk about a paradigm shift, it is certainly arguable that biology urgently needs a sophisticated theory of how symbols work substantially more sophisticated than that implicit in the HGP; Biosemiotics affords a framework in which this might be tried. Indeed, as this paper concludes, there may yet be room for a "Bionoetics", a perspective in which biological explanation can be extended to include cognition in all its forms. Finally, a working sketch of a modeling environment written in LISP, one that shows promise in reflecting the complexities discussed in the paper, is included.

Keywords: Biosemiotics; keyword analysis; syntax; genome; SIRI; GUT; reductionism

1. A SHORT LISP TUTORIAL

Like the computer language LISP, DNA is homoiconic. Thus, a single string can be either program or data, depending on the context. Lisp uses the quote mechanism to achieve this. In particular, LISP normally uses the following syntax;

(function parameters)

This is called a form. LISP has been nicknames "lots of irritating silly parentheses" and it is fair to say that the syntax, involving embedded parentheses to arbitrary recursive depts., is initially off-putting. I ask the reader to be patient, because I believe the pay-off to be worth it.

Very simply, (+ 5 3) will result in 8 being returned by the interpreter symbolized by ">"

So this is a dialogue;

> (+ 5 3)
8

We may wish, for whatever reason, to have (+ 5 3) regarded simply as a list. In that case, we put a "front"
quote in front of it '(+ 5 3).. Now look what happens;
> '(+ 5 3)

(+ 5 3)

So now we have non-coding, "silent' DNA. Alternative splicing in genetics is similar to an alternative prepositional phrase attachment that has been rejected. It is only a minor stretch to think of a "front" quote as modeling this. Short et al (2008) indicate how this could be used by evolution to try different possibilities.

However, LISP is infinitely more subtle than this. It may be the case that we want part of a list evaluated, and the other part left as it is. For this we can use `, or backquote. The use of comma (,) in conjunction with backquote stipulates that everything in the form immediately following the comma is to be evaluated, and the rest is to be left as it is.

Thus, we can construct a list like (print `(the answer is , (+ 5 3)) and when we get this evaluated we see;
> (print `(the answer is , (+ 5 3))

(the answer is (8))

We can get rid of the parenthesis with @
> (print `(the answer is , @(+ 5 3))

(the answer is 8)

Now we have the capacity to model how various parts of a string of nucleotides might be interpreted as program or as data. There is one farther step to take; the issue of how the same string can be now program, now data, and this can be taken on faith unless the reader wants to consult the addendum at the end of this paper.

2 BIOLOGICAL COGNITION, COGNITIVE BIOLOGY

The argument of this paper in a nutshell is the following; the pressure to consider Biology to be largely a computational science in some non-trivial way will intensify in the years immediately ahead. Coupled with this will come an accompanying impetus to reify biological process with statistical analysis as its fundamental touchstone. This has already been attempted in computational analysis of natural language and has largely failed with syntax and semantics now seen as also necessary. What biology may indeed need is an articulated account of how its symbols function, and a modeling environment in which to express this. The following paper makes a gesture in this direction, with code in the appendix, and then considers larger issues of how to consider cognition as part of nature; eventually to turn the tables and ask in what sense cognition - and perhaps computation - is a biological phenomenon.

About a generation ago, it became clear that a "computational paradox" loomed in the cognitive sciences. A similar one obtains in the biology of the early third millennium. The cognitive sciences had witnessed a spectacular series of successes in AI starting from the 1950's equaled in magnitude only by the distressing failure of the early AI systems to scale, or indeed to function in real-world environments. At this point, many of us, including the present author (2003) and Bickhard (2009) following Edelman (Rose, 2003)began to argue for the necessity of biological foundation for a putatively unified "cognitive science". This of course is consonant with the reductionsist drive in science, and led to an eschatological drive in cognitive science

called “eliminative materialism”. This drive needs to be commented on before we proceed further. .

As described in Gleick's biography, Richard Feynman used to erupt volcanically when asked by a journalist for a sound-bite-sized “theory of everything” (GUT). A great deal of physics, as Huxley put it, boils down to ever more precise measurements. It is premature closure to imagine that a few sentences, from however venerable a source, can elucidate anything significant.

There is a deeper and related issue; the status of mathematical physics. The viewpoint taken here is that mathematics is a very compressed and laconic language, arrived at through intense negotiation over millennia between constraints due to our symbolic apparatus, the physical world, and what we can communicate to each other. Consequently, the sin of “psychologism” - that attempt every generation to reduce math to psychological processes – is akin to a labour negotiator pretending that a set of agreements no longer holds, that all power is with the employer or with the union, and the other side must conform.

Yet mathematics is the holy grail of all Reductionism- a set of equations that will explain everything about what we and the world are. Such a set of equations would be, even in principle, subject to constraints due to a set of rules arising from axioms and theorems. It would be complex in the extreme. It would not explain subjectivity, which has a deus ex machina status in QM, or information, which is implicit in nature. It would rest on the creaky philosophical foundations of math.

None of this is really important, because few if any reductions from one discipline to another have been successful in the entire history of science. While chemistry is in principle reducible to physics, the complexity of fields like quantum chemistry indicates that the end is nowhere in sight. Biology seems to require codes, nowhere clear in chemistry. Reductionism, to be successful, requires that one would know precisely the phenomena to be explained, the depth of reduction required, and the terms in which the Reductionism is to take place.

Better perhaps to begin to re-introduce the word “explanation”. This is particularly the case in that there do seem to be distinctions inherent to nature between the “merely” physical, the biological and the conscious – just for starters. What we can salvage from the Reductionist enterprise is the notion of constraints on our explanations. Precisely because the GUT would require tensors of the fourth order, used in GR, so cognitive science is constrained to allow that the brain uses these. In fact, the apparatus of recursions, formal grammars, and tensor calculus that we were using in the late 1990's before the fmri “results” began pouring in seems now to be judicious in the extreme.

With our renovated cognitive science comes an appropriate set of technologies, from HCI (including BCI) to AI to cognitive therapy. That is the most profound benefit from eschewing the eschatological Reductionism of eliminative materialism; we can actually do things with the tools described at the level of finite automata, algorithms and indeed – to some extent – subjective states.

Biology is going through a similar phase to that of the halcyon era in AI, with the $1k human genome now a reality. Skeptics question whether this computational ethos can really capture living systems, with Conrad Waddinton's pithy remark of 50 years ago that DNA bears the same relation to life as the London phone book bears to the social life of London now seeming prescient. The reply, of course, is that the computational ethos works. The argument of this paper is that it will work only up to a point, and that symbolic functioning introduces into Biology the necessity for a new set of concepts in order to avoid making the same mistakes as AI, exemplified by the SIRI system. In particular, molecular biology needs syntax just as much as AI does.

At the risk of making a new set of mistakes, this paper goes considerably further, if tentatively so. For writers like Goodwin (2001), Lewontin and Rose have argued passionately for a wholly new paradigm in Biology. The answer, as before, is that reductionism works, and that there is no indication that a new paradigm is necessary. Yet certain of the "new paradigm" arguments now seem unassailable; living systems function far from thermodynamic equilibrium, species and environment co-evolve, non-linearity is necessary to describe living systems, and teleological reasoning is necessary not just for the mundane purpose of explain what the heart is for, but to describe how the multiplicity of possible proteins is constrained. In fact, Biology may need an anthropic principle of some sort.

The most extreme position taken in this paper argues that biology needs symbolic and cognitive foundations just as much as "cognitive science" needs biological foundations. Alternatively put, an appropriately chosen set of (meta) concepts can encompass both Biology and "cognitive science" in the wild (to use Ed Hutchins phrase). These concepts, coupled with teleolonomy, the notion that living systems act as if they have purpose, result in a perspective called Bionoetics. It is this writer's hope that, having read this paper, the reader will be convinced of the necessity of taking at least a few steps in the direction proposed by the paper.

There are various striking themes in current biology that indicate its massive ambition ;

Mapping ever mere complex biochemical pathways, which often include sophisticated gene expression networks;

Looking at how drugs can be delivered by plotting the mechanical and biochemical nature of the targets with the added impetus of the possibility of using nanoengineering to safely deliver drugs;
Understanding stem cells by factoring in the whole area of epigenetics – and physics, as shown by use of Hooke's law;
(Good Old-fashioned) GOF biology; for example, ever more refined examinations of meiosis and mitosis;
Attempts to look at cancer sub specie viruses, and treat it sometimes qua its physical properties, thus perhaps neglecting the critical issue of aneuploidy;
Much use of biomimicry, particularly in looking at metabolism of cellulose;
Outliers like the aneuploid theory of cancer

To this ambition is coupled relative agnosia about paradigms, not a bad thing if the goal is as pure as explicating a biochemical pathway. Yet this agnosia borders on ignorance (O Nuallain 2008 a); the Keynes remark that anti-intellectuals are unknowingly in the grip of some conservative economist applies a fortiori to biology, which in the absence of theory still has an implicit notion of the gene qua Cartesian homunculus/CPU. The HGP blundered into precisely the same set of errors that natural language processing by computer (nlpbc) had dome a generation before. Sadly, Biology has yet to produce a metatheory/paradigm that will prevent such mistakes recurring, as they just have, with a cost this writer correctly predicted in the tens of billions – at least – to Apple's share price. Yet a few other research and industrial trends can also be pointed out that indicate how working biologists brilliantly make up theory on the fly, as they also show how the legal framework is in a state of flux;

Cancer work on double stranded breaks shows appropriate sophistication in dealing with the interaction between biochemical processes and gene expression, indicating that - given sufficient urgency in the subject matter - common sense prevails (Volcic et al, 2012);

Despite this, the lack of a clear formalism to handle non-homologous recombination – a formalism that a Biosemiotic approach (see below) to gene expression could supply – has made the aneuploidy theory of cancer more refractory than it should be (Pellman, 2007);

The recent Supreme court decisions on the Myriad case makes it urgent that we find a formalism to explicate in full what the informational processes involved in life actually are in a way that does justice to the diverse roles of DNA, RNA, and all other factors;
Provocative work by Pellionisz – who may turn out to be regrettably ahead of his time in genomics as in neuroscience - correctly brings recursion into focus;

It is established at this point even in the popular press (like the NY Times of 2 April 2012) that identical twin studies have indicated that disease markers are relatively inscrutable in the DNA;
The role of viruses in editing DNA may extend even to their effectively creating whole new grammars (see O Nualláin, 2008 a for a tutorial on this) and there are attested cases of endogenous viruses causing speciation in eukaryotes;
Less glaringly, mechanisms like alternative splicing also indicate that our current construction of phylogenetic trees is too simple-minded. Sometimes divergence of evolutionary evidence from morphological and molecular sources leads to ambiguity about how to classify a species (Cohena et al 2005);
The number of mechanisms other than those proposed by Crick's "central dogma", advanced by researchers like Tao Pan is reaching well into the hundreds. Remarkably, it is possible that mai-based cures that like about to be attempted for Hepatitis C (with very promising clinical trials in San Francisco in late 2011) might have been discovered long ago, sans central dogma. The FDA has just given tentative approval to miravirsen

Many years after the epidemic of Kaposi's Sarcoma, the functioning of SOX is till being teased out; an endonuclease in vivo, an exonuclease in vitro, it is selective or not in its destruction of mRNA in a manner that has yet to be resolved (Glaunsinger et al, 2005, now looks naïve). A new formalism is necessary.

It is clear that Biology needs syntax quite as much as natural language processing by computer (nlpbc) and its protagonists need to avoid future SIRI moments in court. Some of the discussion that follows is an exercise in Biosemiotics. In its weak form, this subject studies nature through the lens of signal and symbol theory; in its strong form, it regards Biology as fundamentally a semiotics science. As such, it regards linguistics as a biological subject. Yet, with the exception of Witzany's work (see my reviews in triple-c), Biosemiotics has unfortunately fallen under the spell of Peircean semiology to the point of ignoring what biology needs from it; syntax and semantics in action.
We can go a stage further and, following this author, introduce a new term called "Bionoetics' which looks at Mind manifest in the interaction of organism, population and species with their environment over time. Considered as focusing on a single human subject, this subject becomes cognitive science; out in the field (literally) it subsumes ethology. Biosemiotics and Bionoetics together constitute something approaching a new paradigm. They allow a creative fusion of the human and biological sciences, while adding a new array of tools and metatheory to biology.
In Bionoetics, we consider the principle unit of study as function over time. In Wittgenstein's classic dictum, the meaning of a word is its use in real life; the scenario we're proposing suggests that we should ignore details of specific proteins and so on as

irrelevant when first characterizing a process. For example, sequence similarity is not sufficient to propose functioning mimic epitopes; the pioneering work of Rachel Carson and Thea Coburn demonstrated that often there is no apparent surface similarity – whatever about at the quantum level – between the artificial and natural chemical that is being imitated with tragic consequences. Mimesis in nature involves species often using vastly different genetic and biochemical processes to the species that they are imitating in order to achieve an effect that seems similar to its target audience (Moffett, 1995).

THE SADNESS OF SIRI

The central function envisaged here is a user asking a direction or price question using a cell phone. Speaker-independent digit recognition is already available on many cell phones. The basic vocabulary needed for direction and prices in a specific area comes only to a few hundred words. Many phone apps have been developed to exploit this opportunity; the most high-profile and controversial one is Apple's SIRI.

Apple may have prematurely launched SIRI because of Steve Job's imminent death; the ad campaign for which it is being sued, which features celebrities like John Malkovich, was apparently released very shortly before Jobs' death, and essentially is a rerun of a 1987 Apple video of an ideal for human-computer interaction (hci). Indeed, Malkovich asks his SIRI to tell him jokes. Apple's response to the lawsuit is instructive; they claim SIRI is still in Beta state, and that the plaintiffs had 30 days in which they could return their phones for attested malfunction. In other words, Apple does NOT claim that SIRI works as advertised. The remainder of this section explains why and proposes an alternative; a similar argument will be made later for computational models of gene expression. In particular, it will be argued that the programming language LISP offers several of the facilities needed by SIRI-type systems, and by modeling of gene expression. LISP allows programs and data to have the same form; this is clearly relevant for gene expression. Indeed, the same text in a program can now be a program, now be data. Other advantages are pointed out below. Let's return to the issue of how to do SIRI tasks correctly, and then see where this takes us with the HGP.

The scenario proposed involves a hard-wired speaker-independent vocabulary and context-sensitive uploading of sites particular to the region in which GPS locates the user at a particular moment. This uploading can involve redundant definition of specific words with the phonetic string passed on to Nuance (aka Dragon dictate, which Apple uses for SIRI).

Remarkably, speech will actually make the parser more robust. For example, interrogatives in English about location default to "where" plus "object";

Where is the Metropolitan museum?/Where can I find the Metropolitan museum? Both default to "where" and "Metropolitan museum". The same goes for the equivalents in French (ou), Spanish (Donde), German (wo), and Italian (dove). So the parser needs to hear only "Where" and "Metropolitan museum".

Thus, terms like "where" would be hard-wired whereas "Book of Kells" would be uploaded in the Dublin area, and "Metropolitan museum" in the New York area. There are several advantages in this domain of using speech instead of typed input. First of all are the obvious ones; speed, and ease of use.

The speech interface can simply tag all the other words as noise and exclude them from the parse. Indeed, for these questions, a simple keyword system would suffice and this SIRI could do. That is not the case for more complex questions like,;"How much is it to go from Dublin City Center to the airport by taxi vis a vis public transport?"

For this, we need a much more sophisticated parser – and yet it can be done without recourse to a full syntactic parse. We can use the mechanism variously known as the "semantic grammar" or "Att". This allows implementation of a basic structure;

<interrogative word> <noise1> <object 1 ><noise2><object 2>…..

Here "How much = <interrogative word><noise1> = is it to go from<object 1 > = Dublin City Center etc

The vocabulary envisaged here has the following "cost" words ;English;"What price /How much for" for example;
How much for a pint of Guinness? Or "How much is it to go from Dublin City Center to the airport?"
What does <noise1> <object 1 > ><noise2><object 2> cost?
What does <noise1> <object 1 > ><noise2><object 2> <<noisen> <object n > (etc) cost?

In the 1980's, a set of parsers was produced in LISP (for example, Winston et al 1989) that exploited the attributes of LISP mentioned above, plus macros (see my 2008 paper) to provide an elegant formalism for Q+A systems and syntax in general. Let's now consider the parallels with gene expression yet again.

GENES AND LISP

One central problem has been the attempt to maintain simultaneously

the definition of the gene as the unit of inheritance
the spatial identification of the gene with a string of nucleotides

This essentially corresponds in nlpbc to attempting to perform all the tasks we humans do with language by keywords. That the results of this effort are risible is best attested by the lawsuit against apple for its system SIRI that claimed to "understand' speaker's intent using this method.

We humans use a variety of techniques apart from keyword; syntax, semantics, pragmatics, and indeed leaving information out, the dog that doesn't bark in the night (O Nuallain et al 2007). Similarly, gene expression uses a rich palate of regulons, operators, promoters, transcriptional factors, and so on at the atomic level of strings of nucleotides; at higher logical levels, there are gating mechanisms involving hox genes, syntactical ambiguity involved in alternative splicing, and master regulatory genes. This is all before we start getting into the role of non-dogma mechanisms like retroviruses, and the increasingly active role now seen by the ribosome in selecting amino acids, to take two examples wherein Crick's portentous shibboleth fails to ward off the dark hordes of nescience.

Indeed, if genes are units of inheritance, it is precisely the absence of any sequence of nucleotides that occasionally characterize them. For example, the main distinction between the two finch species scandens and fortis is cultural; it is a song learned from the father without any genetic correlate. So we now must revisit Waddington and note that it is perhaps best to consider inheritance wrt function in the environment as well as changes to the phenotype. For example, melanism, like blonde hair in humans, can be achieved by a variety of mechanisms; field mice may choose to use mc1r or not. Finally, there is increasingly solid evidence of epigenetic inheritance.

Evolutionary theory has been hampered by the fact that the paradigms used have been atomistic; in actual practice, as illustrated in my 2008 paper in Cosmos and History, it is open to evolution to leave an entire alternative sequence of codons inert or not, the better to see how the "hopeful monsters" emerging will functions. The programming language LISP, through the backquote mechanism, allows a facility for modeling this .

Similarly, LISP's object-oriented system (CLOS) allows inheritance so that the multifunctionality of SOX can be easily encompassed. In this paper, we are not going

to get quite so far; what we're going to outline is a simple example of how a "semantic grammar" can elucidate polypeptide structure.

So let's decouple "gene" and "inheritance". In fact, let's expand our study of inheritance to explicit changes to the phenotype, for which we will seek a causal mechanism within the organism, and modulation of how the organism functions in the environment, for which we'll seek an explanation in terms of informational interaction between organism, population and species in an ecosystem over time. It is unlikely that our concept of "gene" qua inheritance will now be useful, as there is quite simply too much going on to make the inevitably atomistic approach implied by "gene" qua inheritance in any way helpful. We went through all this in nlpbc some time ago; rather, we go through it once a decade or so. The goal of this paper is partly to try and preempt that happening in biology.

Progress in nlpbc in general has been made when reflective people were allowed take the reins and point out that, while a complete solution to the "problem" is never going to be available, we can gain a good overall perspective AND develop good technologies if we resist the temptation to overstate our case. Then the field regresses with fads like the statistics scare of the 1990's, which attempted to do away with syntax and semantics altogether, and Apple's recent farcical attempt to sell a new phone by prematurely declaring victory. The analogy with the HGP hype surely gives pause?
At the very least, nlpbc has found out the following;

there are various language tasks, from answering a question to creating a novel, which demand separate consideration;
All these tasks evince different relationships between syntax, semantics, pragmatics, and the lexicon/words themselves as the context is alternately restricted and expanded; the latter to increase informational range and evade reader's tedium by freeing up the range of expression;
There are a range of useful techniques that have been developed; the creator of SIRI apparently did not know that the tasks involved are best addressed by a "transition tree" approach which works on the assumption that the context has been restricted sufficiently for syntax to do some of the heavy lifting in reducing alternative readings normally falling to semantics

Piaget's work (see my 2003 book) was fundamentally epistemology; early in his career, he decided that the essence of knowledge could be gleaned from analysis of the contingent facts of its development. Yet he tried to maintain an objective view of the world; for example, number he saw as an "operational synthesis" between the

operations of seriation and cardination, a synthesis that could come into existence only after ”conservation” status was achieved (ages 5-8 approx). Therefore, even a demonstrated ability to count at a much earlier age – by lower animals as well as humans – was not sufficient to convince Piaget that the full concept of number had been achieved before conservation. As a neo-Kantian, he would have been very receptive to the notion of mathematics as a language capable of mediating subject and object.

Modern theorists like Lakoff have extended this line of argument and called for a thoroughgoing subjectivism of number with all the dangers that immediately suggest. My guess is that Piaget indeed has some critical insights – despite his sloppy experimentation. Mathematics has repeatedly been honed to fit with the external world; it is NOT the case that Riemann naively produced a formalism that Einstein exploited, as Riemann was a very sophisticated thinker who was fully cognizant that his formalisms might have unexpected applicability. The initial English translation of Riemann by Clifford made explicit claims about the applicability of non-Euclidean geometries to the physical world. Secondly, mathematics indeed confounds our reason at times; my compatriot George Berkeley made telling points about how we deal with infinities in differential calculus, and such antinomies are skated over thousands of times a day in math classes worldwide.

It has been a mistake in all parts of linguistics – particularly applied areas like computational linguistics – to assume that the distinctions we refer to as “syntactic” and “semantic” will be reflected in explicit, external dichotomies. It was pointed out by Jerry Hobbs, among others, that at a certain degree of restriction of context, many selectional restrictions are appropriated by the syntax from the semantics, leading to the (horribly named but effective) applied “semantic grammars”

Now we come to the opposite move; that of restricting the context so severely that keywords are deemed sufficient (Wittgenstein's labourer saying ”Slab!”). It may be said that this is precisely what early NLP (like ELIZA) attempted, and more recently a huge Human Genome Project was sold to funding agencies on the same basis that strings of text/DNA would reveal everything.

So what we’re left with - in my humble opinion – is symbolic systems that are tested and evolved over millennia, each with a capacity for recursion (which Piaget failed to acknowledge), and an ability to change the relationships of the component parts (syntax, semantics etc) as context becomes more restricted. Mathematics in particular investigates the space of the rational using this method; it produces quite as many

beautiful wrong theories as correct theories. What is left is a mystery about how we eventually achieve some kind – any kind – of consensus that our models of abstractions like black holes are somehow continuous with the mathematics that we know works in the real world for something like harmonic oscillators. That is a topic for the next paper; for the remainder of this one, we are going to expand the ideas - so far just sketched – about language and gene expression.

2. LANGUAGE AND GENE EXPRESSION – DETAILS

It is now almost universally accepted (Dennett, 1995; Kauffmann, 2000) that terms like "sign", "signified", and "semantics" are appropriate to genetic expression. Moreover, as we've seen, a field called "Biosemiotics" with an associated set of journals has come into being. However, how far can the analogy between genetic expression and natural language production and comprehension actually be pushed? This section first looks at natural language, the best-known symbol system. It considers the various attempts that have been made formally to characterise human language. First, we broaden the context to consider language as one symbol system among others, including the genetic code. We then look at language through the prism of formal language theory before detailing some of the ingenious and thorough formalisms within the fields of formal and computational linguistics. We consider the layers within natural language itself, and that consideration leads to speculation about the role of context. Finally, we return to our original starting-point of genetic expression, and give a prognosis as to the likely progress of this field, as seen from a linguistic point of view. The consequences for biology and medicine are then spelled out. It is obviously vital that genetics learns lessons from the mistakes of another field. Just as the goal of eliciting meaning from parsing of strings of symbols proved infinitely more difficult in natural language than anticipated, so the goal of specifying production of proteins from nucleotide sequences is likely to exercise us for several generations.

2. Natural language (nl) and other symbol-systems

2.1 Layers of language and gene expression

Natural language (nl) seems to share with other human symbol systems like those that have been developed through music and art the following attributes (see O Nualláin, Seán (2003, paper 7));

1. A hierarchical organization, whereby sentences and musical phrases both consist of top-level entities like sentence that can be broken down into subunits like noun phrases to be further analysed into words and so on.
2. Formal complexity of a certain degree (specifically for nl, that of indexed grammars,as we shall see), discussion of which will occupy the next subsection
3. A recursive structure. I state that (this is an example, because within the parentheses is a full sentence). Doug Hofstadter (1979) famously applied this notion, and that of self-reference, to music and art as exemplified by Bach, Goedel, and Escher.
4. Processing within micro-domains called contexts. This shall also constitute a full subsection of this paper, and undoubtedly applies to genetic expression
5. Metaphor. For Noyes in the Highwayman, the moon was a ghostly galleon; for Piet Mondrian, a jumble of boxes could be "Broadway boogie-woogie".
6. Emotional impact
7. The possibility of self-reference. This sentence is false. All Cretans are liars, said the Cretan. Hofstadter (ibid.) argued that Bach died having finally composed a self-referential opus.
8. Ambiguity. The importance of this paper cannot be under-stated. The word "egregious" can mean outstandingly good or bad. Alternative splicing in the genome does seem to indicate exploitation of ambiguity by nature at this level.
9. Systematicity. We should be able to say "Cyrano loves the 'precieuse'" because, regardless of semantics, we have already said that " The 'precieuse' loves Cyrano".
10. Duality of structure (in language, the phonological and syntactic levels) initially established, it is appropriate to predicate of language acoustic, semantic and pragmatic levels. We shall later concern ourselves with precisely how many levels to attribute to gene expression.
11. The notion of a native language, the phonotactic details of which become hardwired at the expense of all subsequent language learning. The distinction between "rue" and "roue" is audible; however, many of us enter a phonetic minefield in trying to distinguish "Caen" , and other soundalikes and sayalikes in French, some of which should not be said in polite company.
12. Creativity. Famously, we can all produce and understand an infinite number of sentences, as what we learn are underlying principles. The repeated statement of this idea underpins some of Chomsky's reputation.

Of the above, 5,6,7, and 9 seem inappropriate for gene expression; the rest are at least up for grabs. It is appropriate to unpack 10 right now. It is always best to illustrate these phenomena with the aid of antinomial entities, which break the rules we are explaining. The German "zwei" breaks phonotactic restrictions in English. I will

currently be breaking a syntactic rule that there is no future progressive in English. Semantic solecisms can be found aplenty in existential judgements about arms in Iraq made by our political elites in the early years of this millennium. Finally, were you to ask me whether I would like to stop here, and I said "yes" without understanding that you might wish to take a break, that would constitute a pragmatic error on my part.

Let us clarify. Phonotactic rules govern combination of phonemes into allowable sound-patterns for each particular language. Syntax deals with allowable combination of words. "Semantics" is a much more complex nexus of concepts. One the one hand, it is about meaning; yet elicitation of meaning often requires recourse to a further level, that of pragmatics. On the other hand, "semantic" formalisms are restatements of linguistic propositions, often in terms of formal logic. So "I got rhythm" may admit of the semantic correlate "Got rhythm (x)". How far does that restatement advance one on the way to external "stuff", to "meaning", to specific proteins? There's the rub. It be all that's necessary, or it may be just one step on the road. In general, if context is severely restricted, then restatement in a semantic formalism may be sufficient. In actual fact, syntax, or indeed bare individual words, may be sufficient for full meaning (specification of proteins) if context is severely restricted. The pretext of some of the exaggerated claims that came from the HGP was that context was always so restricted. It is not; we have generations of work ahead of us looking at how nucleotide sequences differentially produce different proteins depending on the metabolic context.

While the spectacle of Connie Chung explaining on primetime TV in 1989 that the genome specified everything we are and could be now looks like a bad joke, it must be said that many diehards still have an eschatological hope that full nucleotide sequencing will reveal all. For example, Nick Campbell in Nature reviews Genomics in February 2004 states that biological development is readable in a deterministic manner from gene sequences. Similarly Goodman et al (2005) imply an aspiration to hard-core genetic determinism. So the "one gene-one enzyme" paradigm of Beadle et al (1941) lives on.

In a sense, genetically transmitted diseases can show that the Word can be misspelled when it takes flesh. The inherited disease familial dysautonomia arises from a single-nucleotide mutation in a gene called IKBKAP (Ast, 2005). This corresponds to the phonotactic and orthographic levels in human speech and text; "Next" and "nest" are similarly different. Likewise, at a lexical level, Tay-Sachs is caused by a misspelling of nucleotides. At the syntactic level, Bcl-x, which governs cell death, can be alternatively

spliced into Bcl-x (S), a promoter of cell-death, and Bcl-x(L), a suppressor thereof (ibid).

A linguistic analogy to this is the AI classic

The robot saw the hill with the telescope.

So did the robot need ocular aids to see the hill, or was it looking bare-sensored at an observatory? Does "with the telescope" attach to the robot or to the hill? This is structural ambiguity; to be more specific, it is a problem of prepositional phrase attachment, which needs access to the semantic or a deeper level to be resolved (despite the heroic efforts of current research at UC Berkeley by Dan Klein and David Hall to solve this with synta). The nature of this layer has not yet been established for the genome.

Strohman (2000) argues, with respect to the possibility of such a layer, that there exists a mesoscopic layer between the genome and phenotype wherein complex organisational states can exist and where there exist networks of regulatory proteins capable of organising patterns of gene expression and much other emergent cellular behaviour in context-dependent ways. In a later summary, Strohman (2003) refers to classic work by Veech et al. (2001) that established that the rate-limiting enzyme controlling which genes are on or off is a function of the entire mesoscopic system. In particular, different enzymes will be used to do the same tasks in different contexts.

And yes, semantics in natural language has proven almost that messy. In O Nuallain (2003) and passim below, this author contrasts the early Wittgenstein's HGP-like sunniness about the prospect of a neat semantic description for language taken as a whole with his later wholesale rejection of that idea, and his insistence that language could be processed only within microcontexts that he called "Language games". So is the case also for gene expression.

Let us summarize the elements of the proposed analogy between the genome and natural language. Nucleotides, the four letters, spell words, the twenty amino acids. These get linked together into chains of varying lengths, the proteins which can be words (let's allow some recursivity), phrases, or sentences. Occasionally, in natural language, we happen on words, phrases, and "collocations" that are (almost) unambiguous in meaning. Technical terns like "Trilobite", "formaldehyde", "fratricide" and proper names like "Thailand" are examples. Collocations include "media circus", "team effort", "world record", and so on. These are low-hanging fruit

for natural language-processing computer programs. Interestingly, they are almost context-independent, in that they have only one widely-used metaphorical meaning that is distinct from a little-used literal meaning. In the extension of this analogy, the HGP picked up only technical terms, proper nouns, and perhaps those collocations which are radically context-independent. So it should come as no surprise that only 2% of diseases currently admit of a straightforward genetic explanation.

Jacob et al. (1962) focussed over a long period of time on the work of regulatory genes in their concept of "operons". These genes were a key to unlock the actions of a set of other genes, which for example would generate trytophan. Regulatory genes, in turn, were switched on and off by proteins. So genes can be turned on and off to respond to the environment.

As summarized in Carroll (2013) p 501, we get a schema that lends itself gratefully to LISP;

Structural gene – encodes structure of protein eg an enzyme.

Regulatory gene – governs expression of structural gene

Repressor – turns off enzyme production

Operator – acceptor of repressor

Operon – set of Structural genes governed by common repressor and operator and usually involved in a common biochemical process.

Let us now revisit the analogy between genome and langauge , which includes "context" qua the environment. Gene expression now includes feedback loops. We effectively need a new HGP for every context that organisms encounter.

Strohman (opera cit) expands on these points. The expression of the genome is regulated through biochemical mechanisms that sense the bioenergetic state of the cell. In particular, the metabolites NAD and NADH and other synoptic signals represent instant by instant changes in the bioenergetic status of the cell. Changes of metabolites like fatty acids and glucose result in differential gene expression through binding to transcription factors. Strohman wishes to stress the consequences for progress in biology and medicine. Metabolism can be altered by environmental factors like sedentary behavior as by gene mutations like amyloid production. Alterations in

metabolism, in non-syntactic phenomena, are the proximate cause of disease, and cures can be sought without interfering with the genome, or formal language.

Specifically, he argues that gene expression may be regulated by NAD dependent histone deacetylase via epigenetic marking of chromosomal histones as during the life extension resulting from caloric restriction. Again, context affects meaning of a language string in the linguistics analogy. (Let us emphasise that the major problem in nl processing has always been ambiguity).Secondly, he continues that alteration in protein amount or structure, as in dystrophin or amyloid may alter the rate of metabolic reactions resulting in an altered phenotype. Finally, posttranslational modifications like phosphorylation modify proteins. The rather messy picture that is emerging reminds one of NL semantics, and indeed of life itself

Strohman is also attempting an oblique attack at the central dogma of molecular biology; the deterministic, linear, uni-directional, and encapsulated path from DNA to phenotype. Specifically, he wishes to frame the relationship of genotype and phenotype in terms of a complementarity between genetics and dynamics, between the language as a formal system and the context, to use the corresponding linguistic terms. He unpacks and extends Waddington's (1966) notion of the "epigenetic landscape" by proposing linkages and feedback loops between the DNA, phenotype, proteins, environment, and behaviour. Kaufmann (2000) remarks that this nexus of gene and environment must be kept at a very specific state, the "edge of chaos", for maximum creativity to occur. Therefore, behaviors might be genetically assimilated to give Lamarckian effects without violating the central dogma that Waddington desired; alternatively, as in the case of the scandens finch on the Galapagos, birdsong might become entirely learned down the patrilineal line if emergent properties from the nexus of organism and environment decides that this is a more economical way of storing it than imposing it on the genome.

Formal Language theory and NL formalisms.

Derek Jarman's classic short film about Wittgenstein proposes that he was nostalgic for the days of "The ice"; clear, lucid theories of language that he later eschewed, perhaps coincidentally (or not) after living for some time in this author's native island of Ireland. . This is particularly the case as there are deep and beautiful connections between formal language theory, set theory, and computability. The Chomsky hierarchy posits languages of different levels of formal complexity. At the top level, level 0, are languages that can be modelled by non recursively-enumerable sets. The next level, level 1, features languages that are recursively enumerable. Both of these

types of language, and the automata that generate and recognise them, are too powerful to be appropriate models for human language. Level 2 languages are generally also considered too powerful to be adequate models for nl; however, Shieber (1984) adduced evidence that Swiss German included constructions that only formal systems as powerful as level 2 could handle. (Partly because Swiss German, even more than other languages, is a set of disparate dialects, this result has never been fully accepted in the linguistic community). It cannot be proven that any randomly chosen level 2 language can be recognised in finite time; likewise, computational problems at this level cannot have posited of them a solution in finite time. These connections are reminiscent of the unexpected connections between fractals and chaos; a priori, there is no reason to anticipate this type of phenomenon.

The final explicitly recursive level, level 3, is the consensus location for nl. Even Swiss German can be handled by a relatively trivial addition to level 3 called indexed grammars. Level 3 is also called the level of "phrase-structure" grammars. It is illustrated by sentences such as "The mat, on which the cat sat, which Séamus made, is in the hall" where we need agreement between subject and its verb often at a considerable lexical distance. We describe such agreement using the mathematical expression (a^nb^n); Swiss German apparently needs constructions like (a^nb^nc^n). Even a^nb^n constructions cannot be handled with any degree of elegance by level 4 grammars.

Of course, we are uncertain as to where the genetic code lies in this schema. Phenomena like alternative splicing indicate that there is some degree of ambiguity, and thus complexity. Hox genes indicate a degree of hierarchical organization, context-sensitivity as defined in the terms of the Chomsky hierarchy (see O Nuallain 2003) and long-distance dependencies, indicating that perhaps the phenomena Shieber remarked on are present also in the genome. In any case, the HGP treated the Genome as if it was similar to the parody of language apparent in the pattern-matching programs of the 1960's. Lisp programs were programmed in these systems with preprogrammed scripts like "I have problems with my x", to which they would reply "Tell me more about x". It is likely that parsing the genome is infinitely more complex than this.

There has been much superb formal and computational linguistics since the 1960's, though the general problem of parsing any arbitrary sentence in any arbitrary natural language and extracting meaning therefrom remains inviolate. Thus, 100% accurate machine translation will forever remain a pipedream. The original so-called Chomskyan "revolution" which at least had the virtue of adding the rigour of Chomsky's teachers to the area, was attacked with respect to its dichotomisation of "syntax" with a "black box" called "semantics". In particular, the early Chomsky had

difficulty with the manifest difference that could be wrought by a quantifier like "some" in the two sentences "Formal linguistics is nonsense" and "Some formal linguistics is nonsense". Thereafter, at one extreme, formalisms like categorial grammar shifted the burden to the lexicon; therefore, the lexical item's (or gene's) possible syntactic and semantic roles were assumed encoded in it. It is likely that genetics will plumb for a formalism like this, even though others like lexical-functional grammar are manifestly superior for language. Chomsky's later X-bar predilection is in keeping with the MIT tendency to impose more on the lexicon in the manner of categorial grammar.

Compositional semantics demanded that the syntactic components should each generate expression in a semantic formalism like lambda-calculus, which could then be summarized through processes like beta-reduction. The resulting output, corresponding to the "meaning", was initially presumed to require no further parsing despite its opaque nature. However, the formal semantic analysis of a sentence like "Would you like to keep quiet for a moment?" in whatever semantic formalism does not constitute its meaning in any real sense. That requires pragmatics; the perlocutionary impetus of the statement in pragmatic terms is a request for silence, not an expression of the speaker's pleasure.

Yet nl processing (nlp) made massive strides for some time; indeed, it is fair to say that an asymptotoic state was reached in the early 1990's. Syntactic parsing is a textbook affair; the national software registry allows downloading of many useful tools. Type 3 grammars are well represented by systems like the Alvey natural language toolkit. This system maps a huge subset of English onto lambda calculus, a semantic formalism whose origins reveal much about the origins of formal computational theory. A process called beta-reduction makes the resulting output more economical. What then? Well, for each specific application, as is mentioned below, the words to be used must be implemented as separate Lisp objects and taught how to interact with the semantic formalism. We are right back to square one; the omni-pervasiveness of context. Likewise, even the best natural language generator systems like Charon has difficulty in formalizing the interplay of linguistic and conceptual elements. Absurd claims dating back to Schank's (1975) work have not helped and perhaps quickened the tendency within genetics to use what looked like a happy language metaphor. It behooves us to look at context more closely.

CONTEXT

While "culture" has claims to be the most confusing world in the English language, even its meaning will depend on context. "Context" is much talked about, and very

little understood. Even the most naïve PR person knows to complain that his egregious statements are being taken "out of context". We have gotten little distance with this term in decades of nl by computer, despite everyone understanding what it is in folk psychological terms. What then is the prognosis for gene expression in context?

Perhaps context can best be introduced by considering the project of the early Wittgenstein (1922), hinted at above. Wittgenstein sought a grand unified theory (GUT) of language. He argued that what was out there in the world was "Sachverhalten" states of affairs. The world, he argues, "is all that is the case". The Sacverhalten could be decomposed into Tatsachen, simple elements. Language, he argued, consisted of propositions that could be decomposed into atomic propositions. Atomic propositions could be mapped on to simple elements by what he called a "private language", idiosyncratic to everyone (Think of it as the instruction set of a computer). Wittgenstein later repudiated this "Tractatus" idea of language in favour of one that gave a pre-eminent role to context. Language, he argued, should be considered in context; language-games like guessing riddles, playing chess, and telling jokes should be the focus of study. Otherwise, he argued, one ended up asking ridiculous questions in language like "What is a language"? To spell it out, one is already immersed in a language-game as one formulates the question, and one cannot bootstrap oneself up to an objective perspective.

Wittgenstein (1967) presses the attack at this point. Language is like a city, he opined, with warrens in old sections contrasting with the rationally laid-out modern thoroughfares that are analogous to scientific discourse in language; we have seen his compatriot Witzany compare these to introns and exons, respectively. The process of interpretation of the language is going to be vastly different as one goes from one section to the other. Relatively clean types of interpretation that even the HGP could support will work for the modern section, but not for the warrens. Of course, Wittgenstein should not be our main source here. There is plenty of other evidence for the hypothesis that language does not admit of a single method of analysis and, by extension, that we need to be very careful about positing a monolithic type of gene expression.

Even allowing for the Stalinist bias he was forced to impose on his writings, Vygotsky (1962) can responsibly be interpreted as arguing for different evolutionary roots for thought and language. Thought can be discerned in animals' problem-solving; language originates in signalling-systems like birdsong. Piaget's (1972) monumental research oeuvre gave pride of place to "operational knowledge"; conceptual and motor knowledge whose origin is an internalisation of our interaction with the world. For Piaget, language is a form of operational knowledge; Vygotsky, probably more correctly, would emphasise more the autonomy of the linguistic

apparatus. (Indeed, the loss of reputation that Piaget has suffered perhaps can be ascribed to his stubbornness on this point, which made him vulnerable to the Chomskyans as Cognitive Science developed. His most prominent intellectual heir is ironically the renegade Chomskyan George Lakoff, who does not acknowledge Piaget's influence).

Again, the musings of two deceased psychologists may be of less than urgent interest to readers of this book. Yet their speculations gain impetus from a survey of the successes and failures of nlp, and particularly machine translation (MT). Talented researchers like Nirenburg (Nirenburg et al, 1991) and Roger Schank's students (Schank, 1975) implicitly followed the path of the early Wittgenstein, as they sough the Holy Grail of MT; fully automatic translation between human languages by mapping onto a set of language-neutral logical atoms and then mapping from these atoms onto target text. So, for example, we go from "Is fear mé" in Gaelic through a formalism like Schank's conceptual dependency, which claimed that all meanings could be expressed in a set numbering in the low teens (it varied) of logical atoms and we output a Gallic Muddy Waters "Je suis un homme" in French.

Both Nirenburg and the Schankian school later saw the error of their ways, and began to produce effective systems that worked within specific contexts. In particular, Schank borrowed a concept from Piaget via Frederic Bartlett of the "script", a stereotyped set of actions. So there is a script for entering a restaurant, for diplomatic visits, for earthquakes, and so on. Again, the parallel with Wittgenstein's "language-games" is striking. The later Schankian systems (Schank et al, 1975) were surprisingly effective within micro-contexts; yet, as we have described above for the Alvey systems, the Schankian logical primitives have to be mapped to specific words in the micro-context (for example, "open", "ask", and "sit" in the restaurant example) for processing to occur. Nirenburg's altogether more detailed work began to distinguish between generally available semantic distinctions (we always consider whether entities are living or not) to context-dependent ones (is that a cheesy teenage song, or not?). Our own work (Ó Nualláin et al., 1994) confirmed this as we used the Alvey tools to construct visual scenes on a screen on the basis of nl input.

Let us recap. All processing of language, and indeed all cognition, is contextual (Ceci et al, 1994). (Indeed, this author would argue (2008 b) that the cognitive role of self is to prevent information overload by keeping contexts separate; the alternative is mental illness. We narrate to ourselves continually to prevent this). The relationship between syntax and semantics was at least addressed in nl processing by computer (nlp) (Lesmo et al, 1985) and needs to be in genetics. Contexts, at first blush, seem to be idiosyncratic interactions between linguistic and operational knowledge, which require

knowing the precise relationship between the words (the genes) and semantic formalism (the metabolic context) in order for correct processing to occur (in order to predict what proteins will be generated). It is likely indeed that all of this holds for genetics; access to entities like perlocution seems to demand knowledge of oneself as the object of another's wishes, and therefore consciousness, and is not relevant to genetics.

Cognition is also massively hypothesis-driven, and it is likely that every act of gene-expression draws in some way on the experience of the whole organism. We can distinguish in language between acts of context-determination, which don't involve any but generally available semantic primitives like +-alive, and processing within contexts where there is interaction between context-specific primitives and words. It is not unreasonable to expect such phenomena in genetic expression. Yet there is a deeper consideration still in language and context with which we will close this section.

Any language processing act involves evidence from orthographic/phonotactic, lexical, syntactic, semantic/operational, and pragmatic sources of knowledge. The classic "pipeline" model sees information travelling strictly from left to right in this schema. So we assemble sounds into words, parse the words syntactically, add some semantic interpretation, and finally take into account where we are in the discourse, what the interlocutor is expecting of us, and arrive at an interpretation and/or course of action. However, as argued in Jackendoff (1987) and elsewhere, interactions can be a great deal more complicated. The letters "I L Y" in a certain context, that of a marriage proposal a la the famous scene in Anna Karenina, can result in a rigorous business contract with punitive buy-out clauses. For Wittgenstein (1967), a labourer saying "slab" was being quite linguistically sufficient in the circumstances of a construction project; he wants a slab, rather than to discourse on the geology of the material. So we can have connections from the orthographic or lexical straight to the pragmatic if context is sufficiently restricted.

Specifically, context seems to deform the layers of language as it becomes restricted in much the same way that gravity deforms space-time as one approaches the surface of a planet. At a level intermediate between reading the New York Times and I L Y, semantic relations are appropriated by the syntax in "semantic grammars" used for natural language interface to public services. The HGP worked on the assumption that the context was always going to be sufficiently restricted for single words to work, and therein lies its failure. It is a valuable lexicographic tool, and therein lies its success. However, we also need syntax, semantics, discourse pragmatics, and enumeration of contexts if nl is anything to go by.

Barbieri's work (1998, 2002) is a fundamental analysis of biosemiotics and related fields like biosemantics. He makes the point from a semiotics point of view that the distinction between context-independent and context-dependent semantic primitives needs to be emphasised in genetics. It may indeed be the case that the HGP, considered thus, has revealed only the low-hanging fruit; context-dependent primitives. Alternatively, it may be useful to consider the HGP's findings as collocations, or indeed as purely a lexicon; in any case, low-level fruit with context rigorously circumscribed.

So is NL anything to go by? The answer is "at least partly". In NL understanding, we are often concerned with getting the gist of a full story, and that is when the full artillery of linguistic techniques is used. In Gene expression, we are initially concerned with building a full organism, so the analogy is appropriate here. Natural language is also used to elicit a simple answer to a simple question, or specify a protein in genetic terms; such elicitation may be available to the technique known as "semantic grammars" in language, no correlate of which has yet been found in genetics. Can it be the case that combinations of metabolic context and nucleotide sequence recur in ways which are useful to predict the generation of proteins? A first step is obviously to do a human genome syntax project in the way that Ast (2004, 2005) has hinted at in his analysis of alternative splicing. Only then can we see if fatty acids are been taken up by transcription factors in systematic ways that echo the appropriation of semantic roles by syntax in "semantic grammars" (Ó Nualláin, 2003, Pp. 121-126).

Sometimes we understand language in order explicitly to act; we map the language onto a set of Piagetian "schemes" (Schankian "scripts"), routinised sequences of behaviour. Similarly, we read in order to evoke a script that we may or may not act on. Such leisure is not available to the genome; undoubtedly, however, the task of organism construction is a s complex as anything in language (and of course itself generates the elements of the linguistic apparatus).

The prognosis for genetics from analysis of the history of nlp must be, then, that it must start research into the syntax of the genome at a level much deeper than mere introns and exons. While there has been preliminary work on this, it is urgent that we decide issues such as what level of the Chomsky hierarchy the genome lies on. Are there Swiss German type constructions for certain creatures? Could it be the case that the difference between human and chimpanzee brains, to take one celebrated example, is that between Chomsky types? The consensus from Salzberg et al. (1998) is that, whereas the genome is susceptible to description by finite-state automata, protein-folding requires dependencies at long distances, and thus the artillery of context-free grammars. Abe et al. (1997) comment on protein prediction.

It must then try and understand the "semantics", the metabolic context for the particular case that Veech et al (2001) examine. Then comes issues of "discourse structure"; how tasks like building and maintaining the integrity of the organism act as high-level goals affecting the minutiae of protein generation. Finally come issues relating to how these elements come together in specific contexts. There are generations of work ahead, all of which will bear fruit for biology and medicine.

Specifically, the template-matching "one gene, one protein/enzyme" story is childish. When we arrive at a more sophisticated understanding, it will have huge consequences for biology and medicine. Gene expression involves interactions of genome, environment, and emergent characteristics of networks of proteins as Strohman (opera cit.) has rightly argued. In this framework, we can begin to understand how exercise and diet affect metabolism, how metabolism affects gene expression, and how health is maintained as the dynamic set of relations it is.

3. CONCLUSION

This paper is a preliminary foray into theory and metatheory in biology. It begins with the similarity between Apple's flawed SIRI system and the similarly flawed HGP. This leads to a discussion of what a more veridical theory of symbolic functioning in nature would look like, and an exhortation that the modeling be done in LISP. Finally, by way of opening a discussion, an overall perspective called "Bionoetics" is motivated from the paradox that, in our current state of knowledge, both life and mind may simultaneously seem totally accidental and totally inevitable and this will be detailed in the next paper.

APPENDIX

In previous work (2008 a) this author alluded to the use of macros in Lisp to model the action of Hox genes. Here, we can elaborate further and claim – pace Atlan and his colleagues – that there is a case for using macros as a formalism that allows the same string of nucleotides both be program and data.

There is a sketch, as yet unimplemented, of what the code would look like in common Lisp in appendix A; Please see Winston and Horn (1989)for auxiliary code

Appendix
At the top level we have

```
(defmacro define-tree (name-of-tree tree-description)
  `(defun ,name-of-tree (word-list)
     (interpret-tree ',tree-description word-list)))
```

This macro is used here

```
(define-tree interface
  (brnchs
    (count > objects if-end-rtn
                  (db-call `(db-count ,objects)))
    (> enumerate > objects if-end-rtn
                       (db-call `(db-show ,objects)))))
```

and this generates `(defun interface

which is a function like this

```
(defun interface (word-list)
  (interpret-tree
    '(brnchs
       (count > objects if-end-rtn
         (db-call `(db-count ,objects)))
       (> enumerate > objects if-end-rtn
         (db-call `(db-show ,objects))))
    word-list))
```

which is, far as I can see, what the macro is meant to generate
This code

```
(defmacro define-tree (name-of-tree tree-description)
  `(defun ,name-of-tree (word-list)

     (interpret-tree ',tree-description word-list)))
```

allows a function (for example find-oxytocin) to be created that will also use the data structure corresponding to oxytocin to be part of the function itself once it is interpreted as data. The following is a sketch of what the rest of the code looks like –

please note > is an instruction to parse a sub-tree and return a parameter corresponding to a successful traversal of that subtree

Polypeptides
(define-tree phe
(interpret-tree
`(brnchs (uuu)
(uuc))
nucleotides]

(define-tree oxytocin
(interpret-tree
(brnchs (> gly >leu > pru > cycs > asn > gln > lle > tyr > cyn
(if-end-rtn (print 'oxytocin)]

We alternatively can have a separate tree for all polypeptides including oxytocin*

REFERENCES

Bernard L. Cohena,, Agata Weydmann (2005) " Molecular evidence that phoronids are a subtaxon of brachiopods(Brachiopoda: Phoronata) and that genetic divergence of metazoan phylabegan long before the early Cambrian" *Organisms, Diversity & Evolution* 5 (2005) 253–273

Bickhard, M. H. (2009). The Biological Foundations of Cognitive Science. *New Ideas in Psychology* 27, 75–84.

Carroll, SB (2013) *Brave Geniu* NY: Crown

Davies, P. (2008) *The Goldilocks enigma* New York: Mariner

Glaunsinger, BLeonard Chavez, and Don Ganem(2005) "The Exonuclease and Host Shutoff Functions of the SOX Protein of Kaposi's Sarcoma-Associated Herpesvirus Are Genetically Separable". *J Virol.* 2005 June; 79(12): 7396–7401.

Goodwin, B (2001) *How the Leopard Changed Its Spots : The Evolution of Complexity*, Princeton University Press (March 1, 2001)

Mark Moffett (1995) "Why I like jumping spiders". *International Wildlife* May-June 1995

O Nualláin, Seán "The Search for Mind" (Third edition) *Intellect*, 2003.

O Nualláin, Seán and R. Strohman "Genome and natural language" in Witzany (ed.) "Biosemiotics in transdisciplinary context. *Proceedings of Biosemiotics* 2006. Helsinki; Umweb (2007) Pp. 249-260.

O Nualláin, Seán (2008)""Remarks on the foundations of Biology" at "Cosmos and History: special issue on 'What is life?'" Vol 4 Nos 1-2.

O Nualláin, Seán and T. Doris (2011) "Consciousness is cheap" Biosemiotics journal DOI: 10.1007/s12304-011-9136-y

Pellman, D (2007) "Cell biology: Aneuploidy and cancer"Nature 446, 38-39 (1 March 2007) | doi:10.1038/446038a; Published online 28 February 2007

S Rose, S (2003) Lifelines: Life beyond the Gene NY: OUP

Short, S. And L. Holland (2008). "The Evolution of Alternative Splicing in the Pax Family: The View from the Basal Chordate Amphioxus " J Mol Evol. 2008 May 14.

Singh GM, Fortin PD, Koglin A, Walsh CT (2008): Hydroxylation of the aspartyl residue in the phytotoxin syringomycin E: Characterization of two candidate hydroxylases AspH and SyrP in Pseudomonas syringae. *Biochemistry* 2008, **47:**11310-11320

Volcic M, Karl S, Baumann B, Salles D, Daniel P, Fulda S, Wiesmüller L. (2012)"NF-κB regulates DNA double-strand break repair in conjunction with BRCA1-CtIP complexes" Nucleic Acids Res. 2012 Jan;40(1):181-95. Epub 2011 Sep 9.

Winston, P and C. Horn (1989) Lisp Addison Wesley. Code is at
http://groups.engin.umd.umich.edu/CIS/course.des/cis479/winstonlisp.html
(Checked 3/26/12 and there are two slight bugs)