Ribosomes & mRNA Alternative Initiation Codons: Where To Start?

Early Development of Tools For Mapping The Translatome

Daniel Steven Davies

BSc Biochemistry (C700) Third Year (2015-2016) – BIOL3034 Module

Supervisor: Dr Mark J Coldwell

3rd Year Dissertation BIOL3034 2015-16 - Supervisor: Dr Mark Coldwell Daniel Steven Davies

Contents

I. Abstract

Translation of mRNA to proteins is affected by many factors. Tracking effects of Alternative Initiation Codons (AICs) is central to understanding translational variation. Using prior AIC studies and a modified version of Intermine, one can predict AICs and their impacts on translation and so on downstream, modelling the translatome to find out how the effects of genetic modifications from artificially introduced vectors on alternative translational initiation site usage can be modelled in silico. In this dissertation, a manual alignment to establish the mapping methodology for AICs is carried out. The early stages of development of a bioinformatic tool set to handle the effects of genetic changes on Alternative Initiation Codon (AIC) placement and of particular focus here, the effects of these AICs on protein translation and localisation, are outlined, along with contextual analysis and suggestions for further development. The tool set will be called INITIATOR SET, its principle component being InitMine (our tool based on Intermine). This lays foundations for a broader method of predicting translational variations in any scenario, with potential uses in oncology and genetic disease treatments. By creating these tools, an indeterminate number of questions may be answered. One such example to be taken as an initial goal is, “can AICs' effects on protein targeting and localisation be predicted bioinformatically?” The process of software iteration development based on the Intermine template proves more challenging than anticipated, preventing completion within the given deadline. Several ideas for future uses and expansions of this tool set are also laid out.

II. Lay Summary

This study is about mRNA (messenger RNA, which carries a version of the instructions from your DNA out into the cell) and the ways in which the ribosome's (a certain set of proteins that makes other proteins) starting point on that mRNA can be changed whether by good things or bad, or by the body or by people. It is hoped that we can figure out how anything different in the mRNA can change which start point is used and so which kind of protein is made, and use computers to track the changes in a new system called

INITIATOR SET. It might be possible to make more narrowed and careful ways to get rid of problems in mRNA and to be more careful how we read and write our own DNA. When anyone's body's information in DNA is written to mRNA for sending out of the nucleus (middle part) in any cell, that is called transcription. When the mRNA is then used to make proteins by a ribosome, that is called translation. There are several things about DNA and mRNA which can change how they are read.

Computer systems up to now have had to work with very small bits of information at a time rather than the whole body at once. This is the beginning of an effort to make a system that puts the pieces together with less work and time taken by people to do it for each bit of DNA or of mRNA.

This will help us to find new ways to remove many diseases such as cancer (body's own cells gone bad) and diabetes (too much or not enough sugar in the blood because the body is not able to sort it out). Using the different ways of joining up the things proteins do that we find, we will be able to fill in more holes in what we know about our own bodies and about how all the living things with cells like ours are put together and are able to live. This is great not just for making people better from diseases, but also for interesting future machines and ideas that come from thinking about the way life is put together, perhaps even so as to change it and make new kinds of life.

III. Acknowledgements

This project would not be possible without the contributions of the following people:

Jo Cowan, for her previous and current work on Alternative Initiation Codons
Dr Mark Coldwell, for his patient guidance and for organising this project in the context of the University of Southampton, and for his part in Jo Cowan's previous and current works, along with his own work on alternative translation start sites.
Justin Clark-Casey, for introducing me to Intermine just when I needed it.
Three anonymous contributors, for helping to understand and set up Intermine on my server and for keeping my servers up to date more often than I would, in spite of their own

problems at home and elsewhere during this time.

Adam Hartline, for coming to the rescue when dealing with Java.

Mircea Filipescu, Paul Kemp, Richard Graham and others for moral support and lightening the mood.

SoMakeIt for use of their hackerspace and the computer parts they didn't need.

Cathal Garvey and the team at Indie Bio EU for setting an encouraging career waypoint in my future with great scientific enthusiasm.

I am indebted to you all.

IV. Abbreviations

3`	At or in the direction of the Three-Prime end of a DNA or RNA sequence
5`	At or in the direction of the Five-Prime end of a DNA or RNA sequence
A	Adenine
AIC	Alternative Initiation Codon
APC	American Power Conversion corporation (licensed to Schneider Electric in Europe)
C	Cytosine
eIFs	eukaryotic Initiation Factors
eORFs	extended Open Reading Frames
ExTATIC	Extensions and Truncations from Alternative Translation Initiation Codon
FASTA	Fast-All (format for protein and nucleotide codes alike – 'works with any alphabet')
G	Guanine
GNU	GNU's Not Unix (free open source software)
GUESS	GUESS Universal Editing Suite & SDK (or Genetic Unified Editing Suite & SDK)
IDE	Integrated Development Environment (software)

INITIATOR SET	INITIATOR SET Now Identified Through Informatics of Alternative Translational Organisation of RNAs & Sequence Editing Techniques
IRC	Internet Relay Chat
IRES(es)	Internal Ribosomal Entry Site(s)
ITAFs	IRES Transacting Factors
LCD	Liquid Crystal Display
LINUX	Linux Is Not UNIX (free open source operating system)
Ltd	Limited company
LXDE	Lightweight X Desktop Environment
PC	Personal Computer
PCR	Polymerase Chain Reaction
RPF	Ribsomally Protected Fragment (of mRNA in Ribosome Profiling)
SDK	Software Development Kit
T	Thymine
TargetP	Neural network-based algorithm & web portal for determining the presence of localisation sequences in proteins
U	Uracil
UML	Unified Modeling Language
uORFs	upstream Open Reading Frames
UPS	Uninterruptible Power Supply
UTR	'Untranslated' Region (canonically)

V. List of Tables and Figures

Section I: a table of contents for this document.

Section IV: a table of abbreviations used in this document.

Section 7: a table of references used in this document.

Figure 1: a diagram of the relation of the software workflow to the processes of alternative initiation in translation.

Figure 2: a screenshot from Supplementary Table 4 in Libreoffice Calc, with highlighting added by the author, provided originally in the supplementary data for (Cowan *et al*, 2014)

Figure 3: a screenshot from the spreadsheet 'mmc3.xls' in Libreoffice Calc, with

highlighting added by the author, provided originally in the supplementary data for (Ingolia *et al*, 2011).
Figure 4: a screenshot from the spreadsheet 'sd01.xls' in Libreoffice Calc, with highlighting added by the author, provided originally in the supplementary data for (Lee *et al*, 2012).
Figure 5: a screenshot of the FASTA file for ADNP from Ensembl, modified by the author to highlight a particular AIC.
Figure 6: a screenshot of the FASTA file content for Cited2 from Ensembl, modified by the author to highlight the AICs and the GC contents of the eORFs associated with them.
Figure 7: a screenshot of the web interface of InitMine in its current condition, after a search for 'human'.
Figure s1: a screenshot of the results from the search in Figure 7.
Figure s2: a screenshot of an underpopulated search result page for a specific Ensembl ID.
Table s1: a list of components used to construct the server used for InitMine, with prices.

1. Introduction

1.1 A Brief Overview of the Scientific Context

In this section, an outline of the development of our current scientific understanding is provided.

1.1.1 Relevant Cellular Processes

When a eukaryotic cell makes a protein, it transcribes from DNA to mRNA, selecting a certain sequence for this transcription in the first place via epigenetics and then further trimming it via alternative splicing. These elements of transcriptional control shape an mRNA and select its components through alternative splicing and their accessibility via several pathways. By the time a ribosome is bound to an mRNA sequence, this sequence has already been tailored to the conditions in the cell at large. It presents RNA folding motifs such as hairpins and quadruplexes (Wendel *et al*, 2014), Internal Ribosomal Entry Sites (usually viral in origin) (IRESes) and/or alternative initiation codons (AICs) via which

the ribosome, guided by eIFs, must make its final selection of sequence to translate. In 1989, Marilyn Kozak published a paper outlining the understanding at that time of the initiation codons and the contexts they must be in (the 'Kozak Consensus') in order for translation by ribosome to commence for a given gene. The potential for the presence of more than one initiation codon in an mRNA was recognised, but the selectivity between these was put down to the Kozak Consensus and ribosomal scanning (Kozak, 1989). AICs allow the initiation of translation of eORFs (extended Open Reading Frames) and uORFs (upstream Open Reading Frames), which might block translation of the usual ORF for the protein, 'dampen' its likelihood of translation or add an extra polypeptide chain to be translated alongside it. AICS may be simply AUGs in non-canonical locations with more or less Kozak-like contexts or they may be other codons, such as AUC, GUG or pretty much anything other than a stop codon on a scale of decreasing translational likelihood in line with increased difference from the AUG codon and with increasingly non-optimal Kozak contexts. They may be placed in the 5`UTR or within the sequence of the mRNA. Sometimes even the 3`UTR can contain AICs (Coldwell, 2015a). DNA repair is also less easy where uORFs are concerned, so these can be indicators of mutagenic susceptibility. These factors may all be affected by subcellular location-specific proteins and conditions too, allowing a great deal of intrinsic specificity in the code of life for the fine-tuning of protein expression. It is suggested that eIF1 quantities are proportionate to and control the stringency of AIC context-based selection for translation (Ivanov *et al*, 2010), yet ironically eIF1 itself is translated from an initiation codon in a weak Kozak context – this appears to form the major control point for how much eIF1 is produced and by which the stringency of translation initiation of other proteins is decided (Miyasaka *et al*, 2010). Post-translational modifications then further secure the timing and quantity of proteins' expression and their exact destinations, and these modifications are potentially based on the sequences added, removed, exposed or hidden from protein modifying processes by alternative translation (Tatematsu *et al*, 2014).

1.1.2 The Development of Methods of Data Accumulation

Detecting the systems and pathways behind these processes has not been easy, and to a notable degree, is yet incomplete. From the electrochemical blotting techniques used to

detect DNA (Southern blot (Southern, 1975)), RNA (northern blot (Alwine *et al*, 1977)) and which proteins are expressed (western blot (Towbin *et al*, 1979)) in the 1970s, through the invention and improvement of PCR (Saiki *et al*, 1988), and onwards to high-throughput parallel sequencing technologies for DNA, and via Reverse Northern Blotting (Callard *et al*, 1994) and Microarrays, RNA and Ribosome Profiling to detect which sequences in mRNA are actually bound with ribosomes at initiation (Ingolia *et al*, 2009). The basic principle of probing a fraction against the gene of interest has been carried through from northern blotting to modern techniques. Now, with capillary gels, Illumina and similar sequencing techniques and the masses of data they have brought, the focus has shifted somewhat into how to handle the data that these techniques produce.

1.1.3 Big Data and What To Do About Them

Even with what would currently appear to be an only partial ability for most clinical scientists to sift logically through all the data that may be potentially relevant to their studies, we are already in a situation today that sees such techniques as CRISPR and Gene Therapies, which can target specific cells and tissues, being rapidly developed. However, these techniques have their limitations and will continue to whilst there is a lack of connection, completeness and universal accessibility between big biological datasets with which to inform their use and further improvement.

For some time, the reductionist approach of experimental and research scientists to understanding the way life works has driven them to ever-more specific and niche studies and collections of data. The sheer quantities of these data, of the scientific papers about them, and the sheer number of datasets they are found in, each describing distinct properties of particular aspects of the processes on which life is based, require complex, dedicated computational efforts to make meaningful use of these data and creates an entire scientific field in its own right; Bioinformatics.

Bioinformatic analysis is a fast-growing field with wide-ranging applications in biology and biochemistry. A bewildering array of websites, databases and services is developing, presenting a smorgasbord of tools in which to lose oneself, trying to find that one program

that can actually analyse one's data in the way needed. Some of these tools fall out of use through sheer obscurity or lack of funding, whilst others thrive far beyond their original intended purposes as scientists add ever more data to the enormous quantities collected already and bioinformaticians have the task of making the most of computing technologies to deal with these data and find from them some meaningful results.

So far, the development of these tools has been largely ad-hoc, with unification efforts generally leading to websites that curate the individual tools in one place, more than attempting to simplify the processes of their use (Gasteiger *et al*, 2003, Blankenberg *et al*, 2010). Due to the large gaps between tool functions that had so far existed, this was over the last decade or two, probably the best approach to take. However, the number and range of tool functions now available has made it possible to start bridging those gaps in the software and to make systems and methods that join these pieces together in a functional information flow and hopefully, to find ways to simplify them and remove redundancies. This process should also ensure that any given function is more likely to remain available as part of a suite maintained for all its functions and not just the most well known parts.

1.2 Intentions and Rationale of Approach

The intention of this project is to contribute to the methods by which to understand the effects of edits and mutations to genes, on which of potentially several AICs will be used and eventually, other aspects of the transcription and translation processes. The Coldwell Lab at the University of Southampton focuses on identifying these alternative initiation codons and gathering data about them to further understand the codons themselves and how their mRNAs are translated to proteins. This includes working with data from ribosome profiling studies, in which Ribosomally Protected Fragments (RPFs) of an enzymatically digested selection of mRNAs undergoing the earliest stage of translation, is used to identify where ribosomes can start assembling, and thus provides evidence to be used in mapping where AICs might be in relation to the genes they alter the translation of. These RPFs have been compared with predicted AICs from (Lee *et al*, 2012 and Ingolia *et al*, 2011) by (Cowan *et al*, 2014), utilising multiple reads for confidence.

To facilitate AIC mapping, tools such as ExTATIC and FastaBuilder were used, which handle some of the processes for bioinformatic analysis of these codons and their sequence contexts, and for subsequent protein targeting.

1.2.1 Previous Tools

Here are some examples of software tools previously developed, which are part of the work which gave rise to the concept for this project.

- 'ExTATIC' (Jo Cowan et al., Manuscript in preparation) is a macro for Microsoft Excel, to perform a stepwise 3` to 5` 'walk' of the 5`UTR starting from the canonical start site, searching specifically for CUG, AUG and GUG codons. This is suggested by the works of (Ingolia *et al*, 2011) and (Lee *et al*, 2012), to be an incomplete method, since Ingolia's predictions and Lee's experimental evidence show other AICs are in use. An example (AUC) of such is provided in the manual alignment carried out in Section 2.2.1.
- 'FastaBuilder' (Jo Cowan et al., Manuscript in preparation) is an AutoHotKeys macro, designed to run the sequence found by ExTATIC through FASTA formatting, then set online target sequence detection and analysis systems such as SignalP (Emanuelsson *et al*, 2007). Unfortunately, this is susceptible to the failure of these external websites and as such, FastaBuilder currently does not function as it was intended to.

These tools require Microsoft Excel to run, limiting their use to systems on which Microsoft Office is installed or to virtual machines or remote desktops with it installed. Given that many bioinformaticians use Linux (and require it to handle many of the big data sets they deal with on a daily basis), this presents a compatibility and scalability problem. The exposure of these limitations in the existing tool set highlights the need for a new approach.

The above tools, although not necessarily withheld themselves from academic use, are in part based on proprietary software systems and standards and reliant on external

providers, posing a risk from the potential for corporate standard changing whims or collapse, to the longevity of the reproducibility of the studies so far conducted. Open standards are also subject to obsolescence, but the availability of the source code in open source software allows more scope for format conversion and data recovery at a later date.

- (Lee *et al*, 2012) also had a 'Leaky Scanning' detection method, although manually applied to the workflow.

A 'Leaky Scanning Detector' allows for bioinformatic discovery of situations where ribosomes have continued translating, when after a stop codon they have encountered an AIC at a certain distance along the mRNA, such as seen in Lon1 translation (Daras *et al*, 2014). Such detection is desirable for tracing the effects of AICs.

- 'WeakAUG' offers a means to use a neural network based algorithm to predict the translation initiation sites of (specifically) mRNAs with AUG codons in weak Kozak contexts (Tikole & Sankararamakrishnan, 2008).
- 'CENTROIDFOLD' used to be a server offering mRNA structural prediction (Sato *et al*, 2009). Unfortunately it now appears to be offline (as of 12th March 2016).
- 'Mogrify' is an algorithm and web portal for the calculation of the transcription factorial changes needed to transdifferentiate cells arbitrarily between (in theory) any human cell type and any other (Rackham *et al*, 2016).
- GWIPS (Michel *et al*, 2015), UCSC Genome Browser (Speir *et al*, 2016) and Ensembl (Flicek *et al*, 2014) each offer potential systems in which the results of AIC mapping could be displayed and integrated with existing databases.
- The University of Cambridge has developed an adaptable framework for bioinformatic database queries known as 'Intermine'. It is capable of obtaining information from its own databases on its local server/network, and cross-referencing the data between these and external data sources, such as Ensembl, to produce a combined report focussed on the search term. To date, this has been used to track down genetic and epigenetic variants for comparison from among a diverse selection of species, with several species-specific 'mines' created by various groups,

and some adjusted to handle different types of DNA or data. However, there are no complete 'mines' dealing specifically with translation and AICs (Smith *et al*, 2012).

The time and scope provided to this researcher for this project prevent the most comprehensively programmed approach from being pursued in the time given, but that does not preclude improvements from being made with respect to the status quo, and potentially continued with after the official period of research under private funding.

1.3 Foreseeable Applications

This tool set, when complete, is likely to be useful in a variety of ways including, but not limited only to:

- Decision making with regard to mutational significance, i.e. in prognoses, diagnoses, developments of treatments and diet.
- Finding safe gene editing loci with the fewest side effects; which are suitable for the removal of oncogenic or otherwise harmful mutations, the insertion of replacement DNA where sections which are required for normal functions are missing, and other variations of this theme.
- Completing the map of alternatively initiated proteins and their pathways, such that they might be compared and used in conjunction with other 'omic maps for holistic understanding of systems biology and its applications to the treatment and/or removal of diseases.
- An overhaul of the Alternative Initiation Codon discovery and analysis process to make it extensible to a wider range of studies
- As a complementary process alongside such algorithms as Mogrify in refining the targeting and implementation of transdifferentiation of cells
- Providing part of the basis for larger software suites similar to the 'Digital Patient' initiative (Vanessa Díaz *et al*, 2013)

1.4 Foreseeable Impacts

There are many holes yet to be filled in scientific understanding, and many ways to approach each of them. However, it is apparent from the scale of data gathered that the bottlenecks to this understanding lie not in the quantity of data gathered, but rather the ability to glean meaning from amongst such quantities. To that end, bioinformatic tools are understood to be highly impactful, rendering simple and quick that which previously daunted scientists as tedious needle-in-haystack searches. The effect of this on the pace of research, confidence in data, collaboration between global teams and the ambition of new projects can be profound.

It is therefore within reason to expect that a successful and functioning assembly of Initmine and more broadly, INITIATOR SET, will allow a more holistic picture of the interactome to be developed, and of Systems Biology in general. This will also provide a basis from which to further scientifically prove the impacts and presence of 'weak' AICs and particularly, of IRESes in the face of pre-established dogma (Coldwell, 2015b). Mutations will be more traceable from phenotypic traits and symptoms. Genetic editing will be possible to plan in more precise knowledge of its effects with respect to cell types, ages and conditions. Mogrify has already set the stage for the obsolescence of stem cell treatments in several cases, with transdifferentiation within reach as a means to bypass the need to revert cells to a pluripotent state, instead allowing them to be directly changed from one type to another by means of epigenetic reprogramming. INITIATOR SET would allow the 'moonlight' roles and side effects of the transcription factors and their associated gene expression patterns (particularly any AIC-linked epigenetic activities) to be further mapped, along with assisting in the elucidation of any undesirable situations which may occur during transitional states. Such more complete maps of systems biology and 'omics will help to bring a more enriched suite of knowledge and bioinformatic parity to wetware, to such projects as the Digital Patient and The GUESS (Davies *et al*, Unpublished).

The furthering of development of these projects will, if sufficiently publicised, stir but also inform further debate on ethical concerns regarding when it is right to alter the genome and to enhance protein production in those suffering deficiencies.

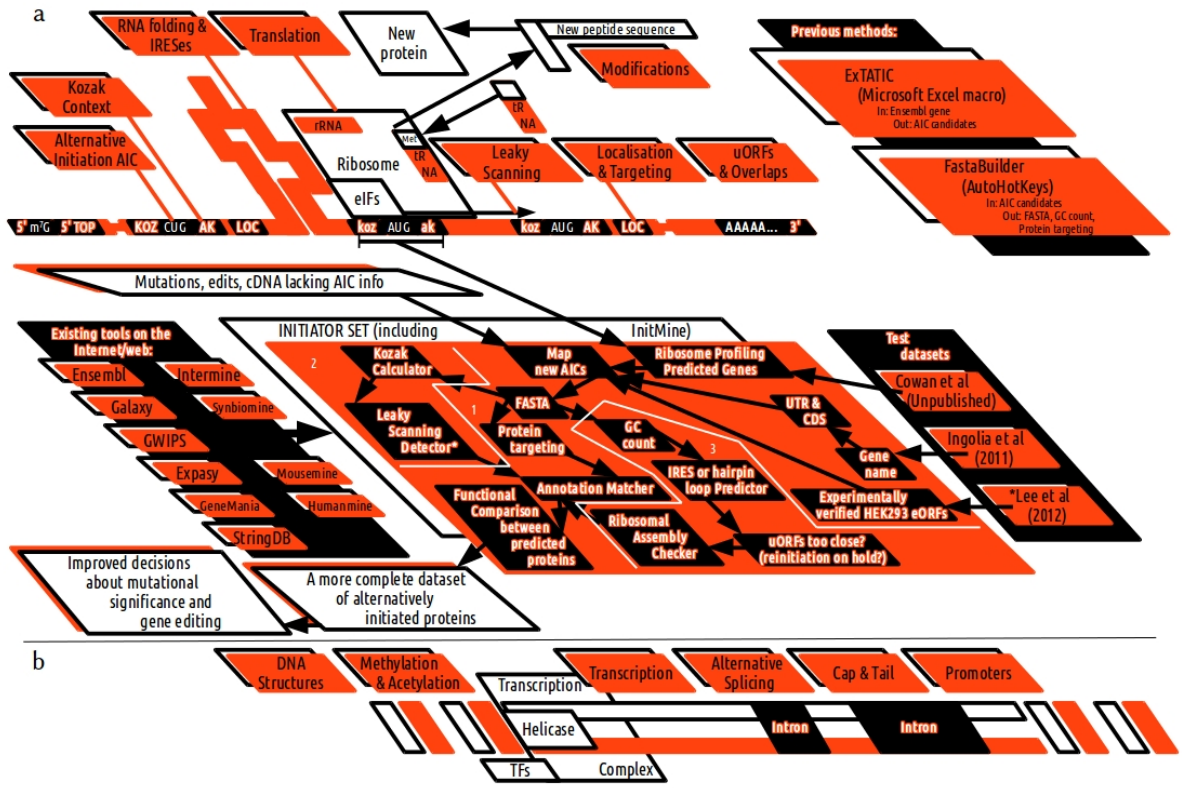


Figure 1: A stylised graphical representation of the potential sources of divergence in a gene's translation (unlikely to all occur together for the same mRNA) (a) and prior to that, transcription (b), and how INITIATOR SET is intended to process these data, particularly those pertaining to AIC translation. Previous methods 'ExTATIC' and 'FastaBuilder' were used in Cowan's paper to process their test dataset. A selection of some examples of the existing tools on the web which may contribute inspiration, part of the functionality or alternatives to parts of INITIATOR SET are shown on the left. From the upper left of (a): mRNAs are often transcribed with AICs included, within the 5`UTR, the canonical start codon and within the canonical ORF. One of the deciding factors as to which of them is selected for a ribosomal assembly to commence, is the Kozak context (previously known as the Kozak Consensus sequence, though it appears there is no consensus about this amongst the scientific community at present). RNA folding (such as hairpin loop structures) and Internal Ribosomal Entry Sites (usually induced by viral activity) (IRESes) affect the ability of the ribosomal complex to assemble at certain locations by merit of convoluting the shape of the mRNA molecule, simultaneously blocking some start sites and rendering others more favourable, including by providing a sort of 'scaffold' structure for the ribosomal assembly to lean on, or to recognise as an eIF substitute. The process of translation itself requires a steady supply of the correct amino acids attached to the correct tRNAs, and the ribosome may under some conditions 'skip' past stop codons or uORFs and continue with a subsequent sequence, in a process known as 'Leaky Scanning'. One potential effect of AIC use is to mediate the exposure of different protein targeting and localisation sequences, which may be buried deep within the canonical form of the protein or may be only present in

an eORF. Some AICs are for sequences offset from the canonical protein coding sequence either by position or reading frame, with the potential to create uORFs, which may overlap the protein coding sequence if present in a different reading frame. The work of Cowan et al was based on earlier works by Ingolia with mouse cells and Lee with HEK293 cells, and took ribosomal profiling data which captured the positions at which ribsomes were being assembled on mRNAs in order to ascertain AIC presence. It is intended that the data from these papers is used to calibrate InitMine to enable it to map new AICs based on data about mutations, edits and cDNA lacking sufficient annotation for manual or macro-based calculations. The Mapped AICs can then be stored in FASTA format for automated analysis by a selection of tools. In the time frame of this project, the focus has been on path '1' – Protein targeting. It is hoped that the other aspects shown can be handled in time too to produce a complete set of annotation which can be used to produce a functional comparison between predicted proteins, leading to a more complete dataset of alternatively intiated proteins and so to improved decisions about mutational significance and the use of gene editing. (b) shows a set of transcription related variables which an additional tool set could then take into account to further expand on this work.

2. Materials & Methods

In this section, the methods used to achieve the current situation are outlined.

2.1 Resource Requirements

With reference to Figure 1, the following resources are required in order to assemble and use INITIATOR SET (insofar as has yet been developed)

2.1.1 Physical resources:

- A computer to function as the server for Intermine, recommended minimum system requirements as per Intermine documentation (Smith *et al*, 2012, see also Supplementary Data). A computer was custom built for this purpose, using components sourced from Aria PC Ltd to specifications outlined in Table s1 (in the Supplementary Data).
- A high speed internet connection with a static IP address to provide a long term stable means of access to the website of the Intermine instance on the server, and also with which to research the scientific papers and data for the project. A suitable

connection and second line were arranged through Andrews & Arnold Ltd for this project.

- A continuous electricity supply compatible with the server's requirements. *The monetary and environmental cost of sourcing this supply necessitates the use of efficient components.*
- A cool, weatherproof location to keep the server in, with ventilation.
- (Desirable) An Uninterruptible Power Supply unit for the server, providing power surge protection and backup batteries to around 4x the idle power usage per server connected to it, to prevent forced shutdowns in the event of power fluctuations or temporary failures, which might affect the integrity of the stored data on the server during read/write operations to/from the data storage media or at any time on the random access memory. This was obtained from Schneider Electric in the form of an APC Smart-UPS X 750VA Rack/Tower LCD 230V (SMX750I), which was also used with other projects and systems at the same time.

2.1.2 Digital resources:

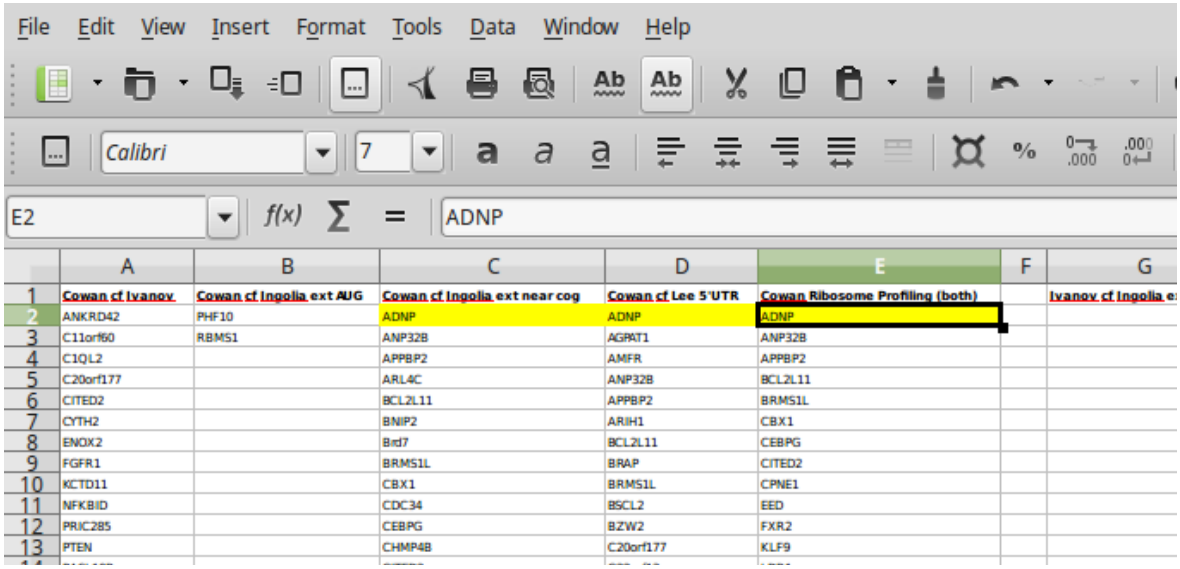
- The Ubuntu GNU/Linux operating system, downloaded from Canonical (free, open source)
- LXDE (to improve computational and electrical resource usage efficiency) (free, open source)
- Oracle Java Runtime Environment (free, closed source)
- Netbeans IDE 7.0.1 (free, mostly-open source)
- Data from various databases online, including Ensembl (free) (Flicek *et al*, 2014)
- Intermine from the University of Cambridge (free, open source) (Smith *et al*, 2012).
- Intermine Documentation (Smith *et al*, 2012) (NB: highly technical).
- A means to contact the developers behind Intermine (in this case, their mailing list via email).
- Optionally, an instant messaging system for real time collaboration with developers and programmers (e.g. Internet Relay Chat (IRC) via the 'Hexchat' client, which is free and open source).

2.2 Input

Outlining the data requirements of the project, starting with a set of manual alignments to establish the data types and workflow.

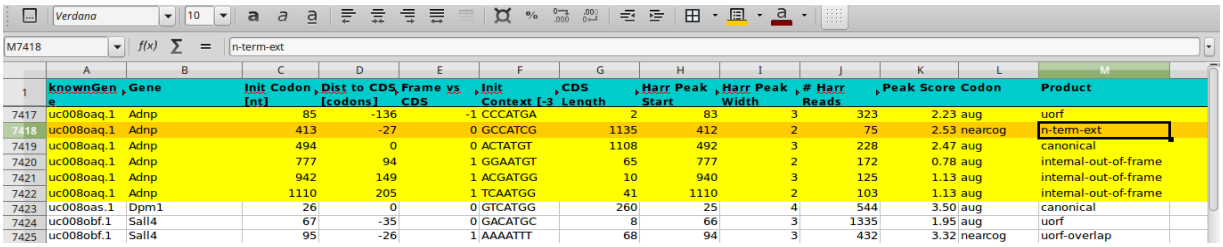
2.2.1 Manual Alignments

As a proof of principle, the initiation codons of one gene were compared manually to find good candidates, based on AIC prediction data from (Cowan *et al*, 2014), (Ingolia *et al*, 2011) and (Lee *et al*, 2012). The gene selected was ADNP. Figures 2-5 are provided to elucidate the process of manual alignment.



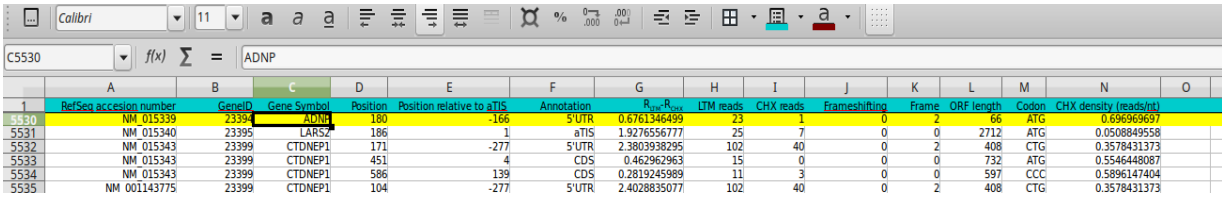
	A	B	C	D	E	F	G
1	Cowan cf Ivanov	Cowan cf Ingolia ext AUG	Cowan cf Ingolia ext near cog	Cowan cf Lee 5'UTR	Cowan Ribosome Profiling (both)		Ivanov cf Ingolia e
2	ANKRD42	PHF10	ADNP	ADNP	ADNP		
3	C11orf60	RBMS1	ANP32B	AGPAT1	ANP32B		
4	C1QL2		APPBP2	AMFR	APPBP2		
5	C20orf177		ARL4C	ANP32B	BCL2L11		
6	CITED2		BCL2L11	APPBP2	BRMS1L		
7	CYTH2		BNIP2	ARIH1	CBX1		
8	ENOX2		Brd7	BCL2L11	CEBPG		
9	FGFR1		BRMS1L	BRAP	CITED2		
10	KCTD11		CBX1	BRMS1L	CPNE1		
11	NFKBID		CDC34	BSCL2	EED		
12	PRIC285		CEBPG	BZW2	FXR2		
13	PTEN		CHMP4B	C20orf177	KLF9		

Figure 2: ADNP was chosen as the first in Cowan et al's list of genes in which AICs had been found via ribosome profiling, which matched both Ingolia et al and Lee et al's datasets.



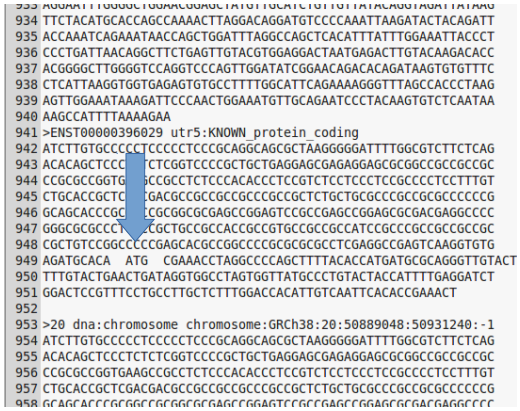
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	knownGene	Gene	Init Codon [nt]	Dist to CDS [codons]	Frame vs CDS	Init Context [-3	CDS Length	Harr Peak Start	Harr Peak Width	# Harr Reads	Peak Score	Codon	Product
7417	uc008oaq.1	Adnp	85	-136	-1	CCCATGA	2	83	3	323	2.23	aug	uorf
7418	uc008oaq.1	Adnp	413	-27	0	GCCATCG	1135	412	2	75	2.53	nearcog	n-term-ext
7419	uc008oaq.1	Adnp	494	0	0	ACTATGT	1108	492	3	228	2.47	aug	canonical
7420	uc008oaq.1	Adnp	777	94	1	GGAATGT	65	777	2	172	0.78	aug	internal-out-of-frame
7421	uc008oaq.1	Adnp	942	149	1	ACGATGG	10	940	3	125	1.13	aug	internal-out-of-frame
7422	uc008oaq.1	Adnp	1110	205	1	TCAATGG	41	1110	2	103	1.13	aug	internal-out-of-frame
7423	uc008oas.1	Dpm1	26	0	0	GTCATGG	260	25	4	544	3.50	aug	canonical
7424	uc008obf.1	Sall4	67	-35	0	GACATGC	8	66	3	1335	1.95	aug	uorf
7425	uc008obf.1	Sall4	95	-26	1	AAAATTT	68	94	3	432	3.32	nearcog	uorf-overlap

Figure 3: The AIC 'ATC' (AUC in RNA terms) at 'position -27' (in the 5`UTR) is the only in-frame AIC detected for this gene by Ingolia et al in the mouse. This gene displays several uORFs.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	RefSeq accesion number	GeneID	Gene Symbol	Position	Position relative to aTIS	Annotation	R_{TIC}:R_{CHX}	LTM reads	CHX reads	Frameshifting	Frame	ORF length	Codon	CHX density (reads/nt)	
5530	NM_015339	23394	ADNP	180	-166	5'UTR	0.6761346499	23	1	0	2	66	ATG	0.696969697	
5531	NM_015340	23395	LARS2	186	1	aTIS	1.9276556777	25	7	0	0	2712	ATG	0.0508849558	
5532	NM_015343	23399	CTDNEP1	171	-277	5'UTR	2.3803938295	102	40	0	2	408	CTG	0.3578431373	
5533	NM_015343	23399	CTDNEP1	451	4	CDS	0.462962963	15	0	0	0	732	ATG	0.5546448087	
5534	NM_015343	23399	CTDNEP1	586	139	CDS	0.2819245989	11	3	0	0	597	CCC	0.5896147404	
5535	NM_001143775	23399	CTDNEP1	104	-277	5'UTR	2.4028835077	102	40	0	2	408	CTG	0.3578431373	

Figure 4: The alternative initiation codon is listed by Lee et al (using HEK293 cells) as being an ATG (AUG) at position -166



```
934 TTCTACATGCACCAGCCAAAACTTAGGACAGGATGTCCCCAAATTAAGATACTACAGATT
935 ACCAAATCAGAAATAACCAGCTGGATTTAGGCCAGCTCACATTTATTTGGAAATTACCCT
936 CCCTGATTAACAGGCTTCTGAGTTGTACGTGGAGGACTAATGAGACTTGTACAAGACACC
937 ACGGGGCTTGGGGTCCAGGTCCCAGTTGGATATCGGAACAGACACAGATAAGTGTGTTTC
938 CTCATTAAGGTGGTGAGAGTGTGCCTTTTGGCATTCAGAAAAGGGTTTAGCCACCCTAAG
939 AGTTGGAAATAAAGATTCCCAACTGGAAATGTTGCAGAATCCCTACAAGTGTCTCAATAA
940 AAGCCATTTTAAAAGAA
941 >ENST00000396029 utr5:KNOWN_protein_coding
942 ATCTTGTGCCCCCCCCCCTCCCGCAGGCAGCGCTAAGGGGGATTTTGGCGTCTTCTCAG
943 ACACAGCTCCC CTCGGTCCCCGCTGCTGAGGAGCGAGAGGAGCGCGGCCGCCGCCGC
944 CCGCGCCGGTG CCGCCTCTCCCACACCCTCCGTCTCCTCCCTCCGCCCCTCCTTTGT
945 CTGCACCGCTC GACGCCGCCGCCGCCCGCCGCTCTGCTGCGCCCGCCGCGCCCCCCG
946 GCAGCACCCGC CGCGGCGCGAGCCGGAGTCCGCCGAGCCGGAGCGCGACGAGGCCCC
947 GGGCGCGCCCT CGCTGCCGCCACCGCCGTGCCGCCGCCATCCGCCCGCCGCCGCCGC
948 CGCTGTCCGGCCCCCGAGCACGCCGGCCCCGCGCGCGCCTCGAGGCCGAGTCAAGGTGTG
949 AGATGCACA ATG CGAAACCTAGGCCCCAGCTTTTACACCATGATGCGCAGGGTTGTACT
950 TTTGTACTGAACTGATAGGTGGCCTAGTGGTTATGCCCTGTACTACCATTTTGAGGATCT
951 GGACTCCGTTTCCTGCCTTGCTCTTTGGACCACATTGTCAATTCACACCGAAACT
952
953 >20 dna:chromosome chromosome:GRCh38:20:50889048:50931240:-1
954 ATCTTGTGCCCCCTCCCCCTCCCGCAGGCAGCGCTAAGGGGGATTTTGGCGTCTTCTCAG
955 ACACAGCTCCCTCTCTCGGTCCCCGCTGCTGAGGAGCGAGAGGAGCGCGGCCGCCGCCGC
956 CCGCGCCGGTGAAGCCGCCTCTCCCACACCCTCCGTCTCCTCCCTCCGCCCCTCCTTTGT
957 CTGCACCGCTCGACGACGCCGCCGCCCGCCGCTCTGCTGCGCCCGCCGCGCCCCCCG
958 GCAGCACCCGCGGCCGCGGCGCGAGCCGGAGTCCGCCGAGCCGGAGCGCGACGAGGCCCC
```

Figure 5: The AIC, then found via a manual count/search of the 5'UTR, is picked out for this image by spacing it apart from the other codons in the human ADNP gene FASTA file as seen here (row 949, indicated by the arrow). It is indeed an ATG (or AUG in the mRNA) corroborated by both sources and by Cowan et al's calculations – or is this a coincidence?

A quick sequence analysis was also carried out using BioEdit, on this occasion no new insights were gleaned. As outlined in the Results section, the manual process is not scalable and the TargetP tool is not sufficient for finding all protein targeting sequences (Emanuelsson *et al*, 2000, 2007).

2.2.2 Software

An interplay between considerations of practicality, time consumption during the longer term project, afterwards for the end users, and acting as a good foundation for future expandability of the project, shaped the outcome of decisions on software to use as a basis for INITIATOR SET.

Intermine was selected due to its versatility, adaptability, open source licensing and development, web-based interface and powerful PostgreSQL & Apache Tomcat database

back-end. Additionally, the author personally maintaining contact with a developer who works on SynBioMine, an iteration of Intermine, allowed for a more informed choice and potential point of contact in case of problems.

During installation and set-up, it became clear that whilst Intermine has these positives, it does suffer in two key areas:

1. it is written in Java, a language which many programmers known to the author online and offline shun owing to its inefficient structure and idiosyncrasies, and
2. its documentation is written with already-advanced bioinformaticians in mind, and does not make for light reading. The documentation mainly serves to show that in order to configure an instance of InitMine, one must alter Java encoded configuration files in a multitude of folder locations and which are often duplicated (but not to be edited) in parallel folders.

These hurdles in conjunction with external unrelated problems proved sufficient to delay progress for a couple of weeks whilst understanding of the structure and configuration of Intermine was established. The FASTA file obtained from Ensembl provided an additional problem by not being integrable into the database build during Intermine setup. After much debugging, the problem was deduced to be the presence of a protein sequence (in Amino Acid alphabet code) amongst the DNA sequences of the file. Removing this sequence fixed that issue, and so it is recommended that future files for input are preprocessed to separate data by type. File and folder structure, permissions and the installation of dependencies also provided snagging points. Different computers with different operating systems and pre-installed software combinations (not to mention, different technical literacy of the person setting it up) will produce different varieties and numbers of such errors and different levels of ease by which they can be resolved.
It is this author's sincere hope that a much less daunting means to configure and deploy an instance of Intermine is created soon, and to that end the author has made efforts to reach out to the developers of Intermine and work with them.

The new iteration of Intermine so created is called Initmine. To then make Initmine into a presentable website such as that seen with the other Intermine iterations, the configuration of Apache Tomcat is required, in our case Tomcat7. This provided further

errors in relation to file permissions and folder structure differences during the installation process, as well as further dependencies on such programs as *tomcat7-admin* and *tomcat7-user,* which are not automatically installed with the rest of *tomcat7*. The user-friendliness and clarity of the installation process was highlighted as a matter of concern with this approach.. Screenshots taken mainly after installation can be seen in Supplementary data.

2.3 InitMine

To set up InitMine, the following steps were taken so far:

- A server fulfilling the minimum system requirements of Intermine to the degree necessary for this implementation, was built (see Appendix 1 for details). The Ubuntu GNU/Linux operating system was installed and configured to use the LXDE desktop environment (when any is needed). This was done using this command which also removes several unnecessary software packages:

```
sudo apt-get autoremove --purge unity unity-common unity-services unity-lens-\* unity-scope-\* unity-webapps-\* gnome-control-center-unity hud libunity-core-6\* libunity-misc4 libunity-webapps\* appmenu-gtk appmenu-gtk3 appmenu-qt\* overlay-scrollbar\* activity-log-manager-control-center firefox-globalmenu thunderbird-globalmenu libufe-xidgetter0 xul-ext-unity xul-ext-webaccounts webaccounts-extension-common xul-ext-websites-integration gnome-control-center gnome-session && sudo rm /usr/lib/thunderbird-addons/extensions/messagingmenu@mozilla.com.xpi && sudo apt-get install lubuntu-desktop
```

- SSH was installed on the server:

```
sudo apt-get install sshd
sudo apt-get install ssh
sudo service sshd start
sudo service ssh-server start
sudo service ssh start
```

- The latest updates were checked for and installed

```
sudo apt-get update
sudo apt-get upgrade
```

- Intermine was downloaded via Git clone as per http://intermine.readthedocs.org/en/latest/git/ and extracted to a working directory created at /srv/intermine
- The directory was changed many times throughout the installation using the 'cd' command, and permissions and ownership of files needed changing in several cases

with the 'chmod' and 'chown' commands.

- A trial run was made using the pre-included 'Malariamine' to test that Intermine can work on this hardware and that all dependencies are installed. http://intermine.readthedocs.org/en/latest/get-started/tutorial/#getting-started
- It became apparent from errors produced, that some more dependencies were needed:

```
sudo apt-get install postgres
sudo apt-get install tomcat
sudo apt-get install tomcatdb
```

- New system user accounts were also needed for Intermine, Postgres and Tomcat to run as system processes.
- A new 'mine' was created as per http://intermine.readthedocs.org/en/latest/get-started/tutorial/#getting-started
- The tutorial instructions then required in-depth and careful reading and understanding in order to appreciate the relationships between the different configuration files and other portions of the programme, such that the right edits could then be made to the configuration files in the right folders, to begin to create a 'mine' suited to the needs of InitMine.
- Our sample data was placed in the appropriate folder to be linked to by the configuration files, particularly project.properties
- The InitMine user account was utilised for configuration file editing and database commands, to prevent permissions issues.
- Edits were made to these various configuration files using Nano and Gedit, for example to allow the InitMine instance to log into and access the Postgres and Tomcat databases, which were separately installed.
- Several attempts were then made to build the 'ant' database. It was discovered during this process that amino acid sequences contained in multi-section FASTA files currently render those files unrecognisable to the system. It was also discovered that incorrect configuration file locations, usernames or passwords and incorrect syntax around these details, will all prevent the database loading

successfully, sometimes without any apparent errors.

- The service processes on the server responsible for Tomcat must also be restarted in full and via the correct method (the shutdown and startup scripts) before each database rebuild attempt or before 'removing and releasing' the webapp.
- It was found that data are only searchable in Intermine in one dataset if another is loaded into which the first can be compared. The documentation was less than clear about these last few points, causing significant delays in this project.

At present, InitMine has access to a minimal database loaded with the FASTA file for ADNP from Ensembl. Its search system is as yet hampered by a lack of optimisation for the types of searches to be done. In the immediate future, the next steps to be carried out will be to create optimal search templates and profiles for automatically gathering the relevant data from the FASTA files and other sources, in order to map AICs and ascertain protein targeting sequences if present. Once this is completed, other data will be added to the database and InitMine can then be applied to the whole list of known AIC-bearing proteins provided by Ingolia *et al*, 2011; Lee *et al*, 2012 and Cowan *et al*, 2014 to map them as a test dataset en masse. After this, the patterns found in the known AIC candidates can be used to seek AICs in other proteins, particularly those of less-canonical sequences (non-Kozak AICs). It will be possible to simply enter a query based on a genetic sequence, and InitMine will compare it with an existing database of AICs, eORFs and protein targeting sequences, producing a list of likely outcomes for the queried sequence, along with links to relevant data about that gene both within the database and on other websites and systems.

3. Results

3.1 Results Drawn From Test Data

3.1.1 Manual Alignment Results

Sequence analysis would seem to suggest that in ADNP, minor mutations are bringing uORFs in and out of frame in direct comparisons between the human and mouse genomes.

Otherwise, would there not be a human AIC at -27? Would there not be a mouse AIC at -166? The functional and sequential relationships between these AICs and their proteins are a matter for investigation. Indeed, how is the canonical start codon of the mouse located relative to the start codon for the human? How similar are mouse and human ADNP? To fully answer these questions, bioinformatic analysis of the AICs of that gene will be needed, but from a visual assessment of the genetic code involved, it is clear that the mouse 5`UTR including the eORF for ADNP is rich in G-C base pairs, whilst the human version is not. Using TargetP (Emanuelsson *et al*, 2000, 2007), it was found that the AIC in ADNP renders both the human and mouse ADNP protein non-secretory, whereas the canonical ADNP variant is secreted.

In reference to Figure 3, having uORFs renders DNA repair difficult on a gene, and thus increases the likelihood of mutations and of more differences between species' versions of the gene. The mouse 5`UTR of ADNP is rich in Gs and Cs, whilst the human 5`UTR is not. A difference between species of the uses of any AICs on this protein in terms of localisation sequences is therefore expected.

Fxr2, another gene identified by all three papers for having a translated AIC, also showed G-C richness of the 5`UTR, including the eORF. TargetP spotted no targeting sequences in either the extended or canonical versions of the protein.

Cited2, found in all three papers as per Fxr2, has two short in-frame eORFs in humans, the longer of these is rich in G-C base pairs, but the shorter is not so much. The canonical ORF is rich in G-C pairs. TargetP again returned a negative result for any differences in targeting between these eORFs and the canonical protein.

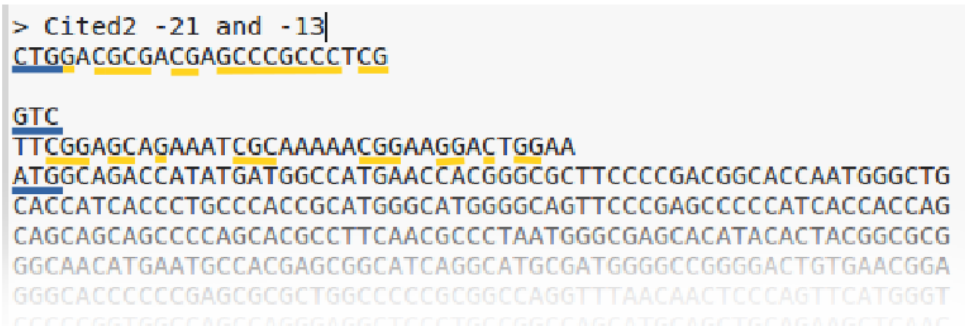


Figure 6: Cited2 exhibits a clear difference between its eORFs' GC content (yellow underlined) by proportional comparison to the size of the eORF overall. Each (e)ORF begins with a different initiation codon (blue underlined).

From these as a basis for further investigation once tools such as InitMine are developed, it can be postulated that GC counts modulate AIC expression, but are not directly responsible for targeting; though their regulatory effect may in some cases be used for selectivity of targeting, targeting is likely not the only reason for AICs to be translated. It should further be noted that TargetP is limited in its target sequence types that it seeks, and a combination of algorithmic tools ought to be used to cover all target sequence types (Emanuelsson *et al*, 2000, 2007).

3.1.2 Test data in InitMine

As yet, no test data AIC automated mapping has been possible owing to delays in development as outlined in the Methods section.

3.2 Results Drawn From New Data

Likewise, no automated mapping of new AICs in InitMine has yet taken place due to the delayed development of InitMine.

4. Discussion

This study has been launched to establish the developmental process of tools to enhance the understanding of AICs to levels of interlinked functionality beyond the current repertoire. The pace of development initially hoped for has yet to be reached, owing to a plethora of setbacks, largely relating to unforeseen deficiencies in documentation clarity for the chosen software base and untimely health issues among those who would otherwise have been able to help with the project more.
In terms of manual alignments, it is important to consider that all data handled were from a set already identified by (Cowan *et al*, 2014) based on (Ingolia *et al*, 2011) and (Lee *et al*, 2012)'s papers, and were used to establish the parameters of AIC mapping for InitMine. These data therefore do not represent a full result set by themselves, and overall were this a project purely aimed at the obtainment of data within a set timeframe, it would be a

failure. However, this project's intended purpose was to commence work on new tools for bioinformatic analysis of AICs, and in that regard it was a success – just not as much of a success in the given time as would have yielded newly processed data in bulk. This paper therefore looks more at how we got here and the intentions of the author for the future of this project.

4.1 Are The Results Interpreted In A Valid Or Reliable Way?

Since this project has a focus on the development of software, there have not been sufficient time or people to carry out mathematical or statistical tests on the data obtained from manual alignments. Interpretations of these manually obtained results are mostly for the purposes of establishing the methods by which InitMine will automatically map AICs. To know whether these interpretations are accurate, it will be necessary to run InitMine with a wide selection of AIC candidates which can be compared with other methods already used. This way, the reliability of the method can also be established.

4.2 How Does InitMine Compare With Prior Tools?

Of course, due to its incompleteness, this is a matter of presenting how InitMine will compare in theory, once completed, based on its expected functionality.
In the long term, InitMine is intended to merge the capabilities of most of the prior tools mentioned in section 1.2.1, with some new functions, as per Figure 1. Tools such as WeakAUG (Tikole & Sankararamakrishnan, 2008) and CENTROIDFOLD (Sato *et al*, 2009) are a good source of inspiration, but do not handle the whole set of potential AICs, thus they cannot simply be included in INITIATOR SET as-is (and in CENTROIDFOLD's case, it appears even accessing a working version may be a tall order). ExTATIC and FastaBuilder go a bit further, but still only cover those AICs with the most frequent observed occurrences, and this therefore leaves out that almost any codon can in theory initiate translation if in a good enough context and not overshadowed by a much more proliferative start codon (Jo Cowan & et al., Manuscript in preparation; Coldwell, 2015b). Whereas these prior tools were 'hard coded' to seek particular AICs, the ability of InitMine

to locate AICs is based on database search templates, which can be modularly adjusted to include the full set of AICs and contexts in queries of a database formed from the FASTA files of the gene(s) in question, or alternatively run as a comparison of database lists. Currently, InitMine is able to display the contents of FASTA files arranged by the Ensembl Identifiers of each sequence listed in them, and it is expected that this limitation is mainly down to the test datasets which have been provided to it so far.

4.3 How Can These Developments Feed Back Into Methodological Improvement?

Previous studies have focussed on building from the original assumption that AUG was the only or main initiation codon, outwards to other codons and potential use of sub-optimal contexts. By using a database type system which starts from the assumption that not only can AUG codons initiate translation in weak contexts (Tikole & Sankararamakrishnan, 2008), but also that any codon can in theory be an initiation codon, except in all likelihood for stop codons (Jo Cowan & et al., Manuscript in preparation; Coldwell, 2015b), it can be then possible to focus on the interactions between any initiation codon and its wider mRNA contexts, translation initiation factors, uORFs and ribosomes, and how mutations and edits to the DNA might affect these in different cell types under different conditions. By allowing for all codons, the metadata about any specific codon can then be adjusted in updates to the system without preventing its position in the wider scope of systems biology being studied via these means. Studies of specific codons and their particular effects in particular proteins will not only continue to be possible through use of INITIATOR SET, but be enhanced by a more expansive selection of tools in one place to apply to these studies. In a context of genetic and biotechnological research in which other areas of the field are already seeing such holistic tools being developed as 'Mogrify' (Rackham *et al*, 2016), the creation of INITIATOR SET and of similar, modularly inter-linkable systems makes a great deal of sense in order to work towards complete modelling of the cell in all its types and situations, and beyond that, the entire body.

4.4 Additional Factors To Consider

It is possible that some AICs are only utilised during specific stages of development, e.g. for embryonic forms and uses of proteins. Haemoglobin provides an example of differential protein morphology in embryonic stages (Alberts, 2015). Some AICs are only utilised in certain tissue or cell types, which can be relevant to many avenues of study, including the development of genetically modified 'bioreactors' for protein mass production, for example from silkworms (Tatematsu *et al*, 2014).

4.5 Further Research Using This System In Its Current Iteration

The main avenues of research for which the current iteration of InitMine will be useful, are into further improvements to bioinformatic systems and their documentation, including that of Intermine itself. With additional data, it will be possible to compare AIC-bearing sequences with one another, and to use InitMine as a portal to other sites and systems, aggregating useful search information. To proceed with protein targeting informatics in relation to AICs, it will be necessary for a template query to be designed and additional data sources to be added to InitMine, these are works in progress now due to be carried out in the weeks after the deadline of this dissertation. A follow-up document detailing the fuller capabilities of future, more complete iterations of InitMine will be produced when the iterations to which they will refer exist, under the auspices of Vulpine Designs Unlimited, potentially in association with the University of Southampton's Coldwell Lab (Davies *et al*, Unpublished).

5. Conclusions

In principle, the plausibility that the method so far followed will lead to an answer to the question of whether we can predict the effects of AICs on protein targeting and localisation sequences, has not been disproven. However, it has also not yet been proven by this method, beyond the outcomes of manual alignments, which have provided the basis for both speculation and the bioinformatic directions in which the system should look. The software certainly bears the appearance and logical capabilities from a technical

standpoint, to produce the desired outcome.

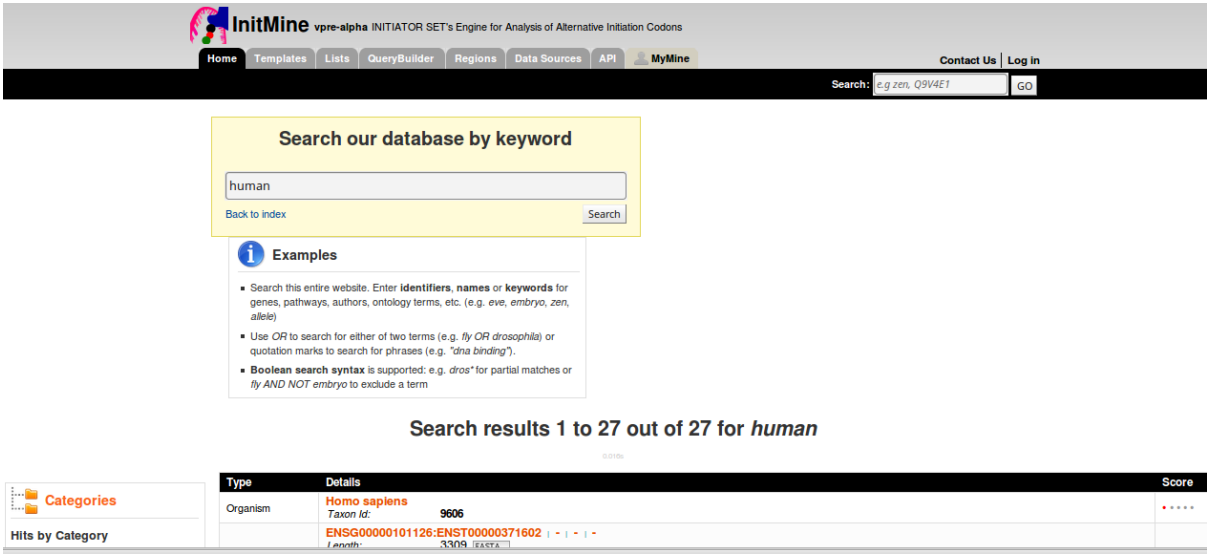


Figure 7: InitMine in its current condition (as of 14th March 2016) is able to perform searches on the very limited set of data it so far has, but is yet to be programmed with specific search templates.

In terms of completion within the deadline of a university dissertation, this project can be considered a failure, although to do so is to not see it in its grander intended context, in which the emphasis is not on the time it will take but on the quality of the end result. Sometimes work and research do not yield results within the time given, and on this occasion it is expected that results will be seen during the continuation of this work outside of the university degree course.

6. Future Work To Improve On This System

So, explanations of misalignments between progress and deadlines aside, here are the main ideas for the future improvements which, once the initial stages are complete, could be made in order to build INITIATOR SET up to be a seriously useful bioinformatic tool set. These form the software development roadmap.

6.1 Kozak Context Calculation

Having already been done on a small scale using a 'neural network' based system (Tikole &

Sankararamakrishnan, 2008), Kozak Context calculation is certainly feasible as a tool for INITIATOR SET (for its position in the workflow, see Figure 1). It will allow the Kozak contexts, that is, the compatibility of the upstream and downstream nucleotides (particularly +4 and -3) from the start site with eIF1, to be compared between AICs and between Before and After mutations or edits, between species, etcetera. Unlike the WeakAUG tool, however, this one will be intended to work equally well on Kozak calculations for all possible initiation codons, taking into account the bias for or against usage introduced by the codon itself and whether any 'stronger' initiation codons are present (including quantifying by how much they are stronger).

6.2 RNA Folding & IRESes

RNA is a flexible polynucleotide chain. Its numerous folds and conformations are essential to complex eukaryotic regulation of protein production.

6.2.1 mRNA Folding Motifs & GC Counts

Particular mRNA sequences are liable to folding and self-annealment. Since mRNA is single stranded, the exposure of hydrophobic bases to the cytoplasm is not very stable as a configuration. Uracil tends to have a preventative impact on the ability of mRNA to self-anneal, but areas rich in Guanine and Cytosine will tend more readily to fold on themselves and create hairpin loops and other structures, such as RNA G-Quadruplexes (Wendel *et al*, 2014). By counting the GC content of mRNA sequences, the regions of that mRNA most likely to fold can be ascertained, and compared with the background cellular chemical conditions and the presence of fold-inhibiting factors. A GC Count Analyser is therefore a valuable tool to create for more complete understanding of mRNA folding and so the availability of mRNAs for translation at a given time and cellular location.

6.2.2 IRESes

Internal Ribosomal Entry Sites (IRESes) can be formed from a variety of complex structural

motifs in mRNA, which arrange the mRNA (usually with a group of proteins known as ITAFs, or IRES Transacting Factors), such as it mimics a collection of Initiation Factors, effectively fooling a ribosome into assembly on that spot. Often the result of viral genetic insertions, IRESes tend to be activated in times of stress, heat shock and other extremes. (King *et al*, 2010) To detect and calculate the effects of IRESes will be a complicated undertaking, owing to their wide variability, but it is hoped a tool can be developed with the ability to distinguish them from other motifs and from plain unfolded mRNA, thus to predict IRESes. Machine pattern learning algorithms may be of use here.

6.3 Ribosomal Assembly Checking

For translation to take place, not only must the right conditions be present in the mRNA itself, but so too must there be a balanced supply of eIFs and other ribosomal components, such as rRNA. Some cellular circumstances and protein conditions might prevent the actual formation of ribosomes, or might change how quickly they can form, thus affecting the ability of a cell or part of it to translate proteins. As Figure 1 shows, this tool is not expected to be included for some time, but it would be very useful to quantify such differences in ribosomal availability and assembly, so as to establish the upper bounds of protein production possible at a given time in a cell or part thereof. In neurons, for example, this could be a major factor in synaptic protein turnover, affecting the rate of growth and signal propagation at long distances from the soma, with a restricted supply of ribosomes and their components available in these distant peripheries. Understanding these processes could be key to advancements in understanding and inducing healing and cellular regenerative medicines.

6.4 Leaky Scanning Detection

As seen in (Lee *et al*, 2012), it is possible to detect instances of ribosomes performing 'leaky scanning' whereupon one initiation codon is missed in order to translate another. This is linked to the activity and concentration of eIF1, and is likely a secondary effect of Kozak context and the other effectors of AICs and mRNA folding. It is desirable to include a

means to detect leaky scanning, to overlap and compare data from this with the outputs of other tools.

6.5 uORFs

uORFs represent a large proportion of AIC initiated sequences. Rather than being an extension to the canonical ORF, these are separate segments of mRNA which can be either offset altogether from the protein sequence, or in an overlapping or offset reading frame, using the same nucleotides to form different codons. Their usage can affect the availability of mRNAs and ribosomes for canonical protein translation; in some cases this appears to be a mechanism specifically for the purpose of preventing too much of a protein being translated. If their usage can be predicted based on the circumstances of other mediating factors and on their locations within the mRNA, another contributing factor to overall translation can be modelled.

6.6 Matching to Experimental data

At this stage, double checking that the results of processing AIC usage data match with experimental data will be very wise, allowing for the quantification of any remaining impacts on translation yet to be modelled or calculated. Datasets from ribosome profiling studies can be compared with the output of the combined tools of INITIATOR SET to identify any further discrepancies.

6.7 Function Prediction

From these data, it ought then be possible to predict the potential functions of a protein, based on its motifs, domains and homologies. A computational method to do this based on known protein folds and patterns will initially not be very accurate, but over time could be improved with pattern learning algorithms and machine learning techniques. This is considered a longer term desirable addition to INITIATOR SET, to utilise computational improvements yet to be made available. In some ways this will be an expansion on

targeting and localisation prediction, but with far reaching implications if successfully created.

6.8 Alternative Splicing Prediction

Directly after transcription, an mRNA molecule may be spliced to different lengths and leave different AICs available for translation on different occasions in different cell types. The spliceosome is in a sense, the bridge between the transcriptome and the translatome. It will be very useful to understand the effects of alternative splicing on the sequence of the mRNAs produced and what conditions lead to which splices. Comparisons of samples from large datasets will enable the understanding of patterns in the splicing of pre-mRNAs into mRNAs, and from this and other researchers' prior works, a tool for alternative splicing prediction can be created which can feed its output directly into InitMine for further checking.

6.9 Promoter Locations

To fully trace the likelihood of transcription occurring in the first place of an mRNA or a particular variant thereof, it is important to understand how the locations of promoters in relation to these genes affects the efficacy of the transcription complex's assembly (mention: Figure 1, section b). It is likely that effects of promoter location are often indirect or loose, nonetheless it seems to be unlikely that there is no effect whatsoever of a large difference in the sharpness of DNA curvature caused by differing promoter locations. Optimal positions are likely to exist, which will be affected by a variety of factors. As this is a fairly nebulous concept to trace, it is unlikely to be modelled in the near future and its inclusion here is to indicate completeness of consideration of potential factors.

6.9.1 Intermine

In tracing the locations of promoters and their effects, the use of Intermine may again be required. Its database comparison methods allow the handling of large datasets and

elucidation of data relationships (Smith *et al*, 2012).

6.10 Cap & Tail

Often overlooked, the m7G cap and Poly-A tail of mRNA are added in different ways and in the case of the Poly-A tail, trimmed to different lengths. The effects of these on the longevity of mRNA and on the potential for a ribosome to be able to translate a short mRNA before binding to the ER translocon, need to be accounted for when considering protein targeting and the overall quantity of a protein likely to be produced before the mRNA is degraded. Enzymes which modify the cap and tail are themselves regulated, so to fully explore the causes of these variations, these pathways need to be followed upstream to identify how cellular conditions and mutations affect the capping and tailing of mRNAs.

6.11 Transcriptomic Analysis

As already alluded to in sections 6.8 and 6.9, and to some lesser degree in 6.10, mapping and analysing the content of the transcriptome, as the source of mRNA polynucleotides, is of great use in determining the overall chain of events and effects between DNA and proteins.

6.11.1 Transcription Factors

Transcription factors are involved in the assembly and regulation of the transcription complex. Being regulated in various ways themselves, they have direct and indirect impacts on the quantity and selection of genes transcribed. They may affect (for example) the tightness of histone winding, the availability of promoters or the phosphorylation of other transcription factors. They may be affected by all sorts of intracellular pathways, some of which are themselves externally regulated. They may also be co-opted by viruses. A system able to translate a diverse range of cellular circumstances and inputs into a set of data regarding the relevance of these to transcription factors, and then from this to determine the effects on genes transcribed, is a very broad and long term goal which would

contribute greatly to the completeness of modelling the factors impacting protein production.

6.11.2 The Transcription Complex

The transcription complex itself, with its promoter binding elements and its helicases, polymerases and other factors, is a complicated set of proteins which must correctly copy many thousands of bases of DNA to mRNA at a time. The potential for mistakes is ever-present; however small the chance is, the vastness of the genome renders errors a regular occurrence. Identifying what parts of the transcription complex are error prone and how the effects of mutations and edits in transcription factors are passed on to other component proteins and so into the mRNA, will further assist in covering every aspect of protein production.

6.12 Rewriting The Software

Naturally, by the time INITIATOR SET and a complementary set of transcription related tools are developed, much more efficient computational approaches and programming techniques will doubtlessly be identified, and like all good bioinformatic systems, they will sooner or later be due for a good overhaul. Perhaps by planning this into the roadmap many years in advance, it might be possible to avoid the fate that befell CENTROIDFOLD (Sato *et al*, 2009) and several other tools which either became obsolete or ran out of funding. Of course, funding itself is a whole other can of proverbial worms deliberately not covered in this document.

6.12.1 Existing Framework & Workflow Tools

Creating a more standardised workflow based on a system such as UML (Yan, 2010; Booch *et al*, 2005) is likely to bring the benefit of a stable and logical plan for long term bioinformatic software development, allowing the project to more readily be continued independently of the availability of individual contributors. Sharing the code in an open

source fashion via such tools as GitLab after its initial framework is established will further expand the potential for voluntary contributions.

6.12.2 Creating A Universal Suite

This is actually not the first launch of any attempt to unify a holistic approach to the modelling of cellular systems biology and thus contribute to the modelling of organisms in full (Vanessa Díaz *et al*, 2013; Yu *et al*, 2013). The GUESS will, if and when completed, utilise INITIATOR SET as one of several adaptable modules undergoing continual improvement (Davies *et al*, Unpublished). The ability to trace as many effects of mutations or edits made deliberately or by viral activity as possible, is essential to a future in which genetic diseases are to be eliminated by humanity taking the reins of its own evolution. A slow start and a steep learning curve this may be for now, but many of the greatest achievements of science often start this way, and whilst delusions of grandeur are not being entertained by this author amid all these setbacks, a little bit of hope is.

7. References

Alberts B (2015) Molecular biology of the cell Sixth edition. New York, NY: Garland Science, Taylor and Francis Group

Alwine JC, Kemp DJ & Stark GR (1977) Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Natl. Acad. Sci.* **74:** 5350–5354

Blankenberg D, Kuster GV, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A & Taylor J (2010) Galaxy: A Web-Based Genome Analysis Tool for Experimentalists. In *Current Protocols in Molecular Biology*, Ausubel FM Brent R Kingston RE Moore DD Seidman JG Smith JA & Struhl K (eds) Hoboken, NJ, USA: John Wiley & Sons, Inc. Available at: http://doi.wiley.com/10.1002/0471142727.mb1910s89 [Accessed November 5, 2015]

Booch G, Rumbaugh J & Jacobson I (2005) The unified modeling language user guide 2nd ed. Upper Saddle River, NJ: Addison-Wesley

Callard D, Lescure B & Mazzolini L (1994) A method for the elimination of false positives generated by the mRNA differential display technique. *BioTechniques* **16:** 1096–1097, 1100–1103

Coldwell MJ (2015a) BIOL3015 MJC Translational Control Lecture 1.

Coldwell MJ (2015b) BIOL3015 MJC Translational Control Lecture 3.

Cowan JL, Perry LS, Edwards RJ, Damerall D, Roworth AP, Johnston HE & Coldwell MJ (2014) Identification of non-AUG initiated, N-terminally extended open reading frames in human genes by an experimentally- informed bioinformatics workflow. *Nucleic Acids Res.* **Unpublished Draft:**

Daras G, Rigas S, Tsitsekian D, Zur H, Tuller T & Hatzopoulos P (2014) Alternative Transcription Initiation and the AUG Context Configuration Control Dual-Organellar Targeting and Functional Competence of Arabidopsis Lon1 Protease. *Mol. Plant* **7:** 989–1005

Davies D, Filipescu M, Hartline A, Anonymous & Anonymous 2 (Unpublished) The GUESS: The GUESS Universal Editing Suite & SDK Cyberspace: Vulpine Designs Unlimited Available at: http://www.vulpinedesigns.co.uk

Emanuelsson O, Brunak S, von Heijne G & Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* **2:** 953–971

Emanuelsson O, Nielsen H, Brunak S & von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300:** 1005–1016

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Giron CG, Gordon L, Hourlier T, Hunt S, Johnson N, Juettemann T, Kahari AK, Keenan S, Kulesha E, et al (2014) Ensembl 2014. *Nucleic Acids Res.* **42:** D749–D755

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD & Bairoch A (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* **31:** 3784–3788

Ingolia NT, Ghaemmaghami S, Newman JRS & Weissman JS (2009) Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* **324:** 218–223

Ingolia NT, Lareau LF & Weissman JS (2011) Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell* **147:** 789–802

Ivanov IP, Loughran G, Sachs MS & Atkins JF (2010) Initiation context modulates autoregulation of eukaryotic translation initiation factor 1 (eIF1). *Proc. Natl. Acad. Sci.* **107:** 18056–18060

Jo Cowan & et al. (Manuscript in preparation) Untitled. *Unpublished*

King HA, Cobbold LC & Willis AE (2010) The role of IRES *trans* -acting factors in

regulating translation initiation. *Biochem. Soc. Trans.* **38:** 1581–1586

Kozak M (1989) Context effects and inefficient initiation at non-AUG codons in eucaryotic cell-free translation systems. *Mol. Cell. Biol.* **9:** 5073–5080

Lee S, Liu B, Lee S, Huang S-X, Shen B & Qian S-B (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci.* **109:** E2424–E2432

Michel AM, Ahern AM, Donohue CA & Baranov PV (2015) GWIPS-viz as a tool for exploring ribosome profiling evidence supporting the synthesis of alternative proteoforms. *PROTEOMICS* **15:** 2410–2416

Miyasaka H, Endo S & Shimizu H (2010) Eukaryotic translation initiation factor 1 (eIF1), the inspector of good AUG context for translation initiation, has an extremely bad AUG context. *J. Biosci. Bioeng.* **109:** 635–637

Rackham OJL, Firas J, Fang H, Oates ME, Holmes ML, Knaupp AS, The FANTOM Consortium, Suzuki H, Nefzger CM, Daub CO, Shin JW, Petretto E, Forrest ARR, Hayashizaki Y, Polo JM & Gough J (2016) A predictive computational framework for direct reprogramming between human cell types. *Nat Genet* **48:** 331–335

Saiki R, Gelfand D, Stoffel S, Scharf S, Higuchi R, Horn G, Mullis K & Erlich H (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239:** 487–491

Sato K, Hamada M, Asai K & Mituyama T (2009) CENTROIDFOLD: a web server for RNA secondary structure prediction. *Nucleic Acids Res.* **37:** W277–W280

Smith RN, Aleksic J, Butano D, Carr A, Contrino S, Hu F, Lyne M, Lyne R, Kalderimis A, Rutherford K, Stepan R, Sullivan J, Wakeling M, Watkins X & Micklem G (2012) InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics* **28:** 3163–3165

Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* **98:** 503–517

Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, Lee BT, Learned K, Karolchik D, Hinrichs AS, Heitner S, Harte RA, Haeussler M, Guruvadoo L, Fujita PA, Eisenhart C, Diekhans M, Clawson H, Casper J, Barber GP, et al (2016) The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.* **44:** D717–D725

Tatematsu K, Uchino K, Sezutsu H & Tamura T (2014) Effect of ATG initiation codon context motifs on the efficiency of translation of mRNA derived from exogenous genes in the transgenic silkworm, Bombyx mori. *SpringerPlus* **3:** 136

Tikole S & Sankararamakrishnan R (2008) Prediction of translation initiation sites in human mRNA sequences with AUG start codon in weak Kozak context: A neural network approach. *Biochem. Biophys. Res. Commun.* **369:** 1166–1168

Towbin H, Staehelin T & Gordon J (1979) Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc. Natl. Acad. Sci.* **76:** 4350–4354

Vanessa Díaz, Marco Viceconti, Veli Stroetmann & Dipak Kalra (2013) Digital Patient Roadmap European Union: DISCIPULUS Available at: http://www.digital-patient.net/files/DP-Roadmap_FINAL_N.pdf [Accessed November 29, 2015]

Wendel H, Singh K, Wolfe A, Zhong Y, Drewe P, Porco J, Pelletier J & Rätsch G (2014) 558 RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. *Eur. J. Cancer* **50:** 181

Yan Q (2010) Bioinformatics for Transporter Pharmacogenomics and Systems Biology: Data Integration and Modeling with UML. In *Membrane Transporters in Drug Discovery and Development*, Yan Q (ed) pp 23–45. Totowa, NJ: Humana Press Available at: http://link.springer.com/10.1007/978-1-60761-700-6_2 [Accessed March 14, 2016]

Yu SJ, Tung TQ, Park J, Lim J & Yoo J (2013) A unified biological modeling and simulation system for analyzing biological reaction networks. *J. Korean Phys. Soc.* **63:** 2247–2254

S. Supplementary Data:

https://intermine.readthedocs.org/en/latest/get-started/tutorial/

is the URL from which the documentation of Intermine can be accessed.

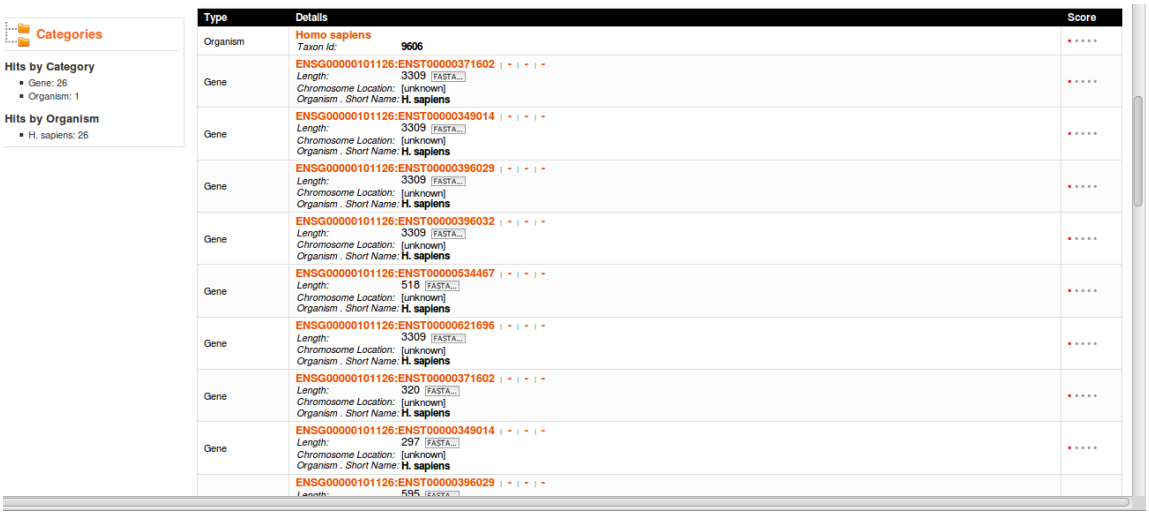


Figure s1: The results from the search pictured in Figure 7. Note that they are currently all listed by the rather user-unfriendly Ensembl ID codes.

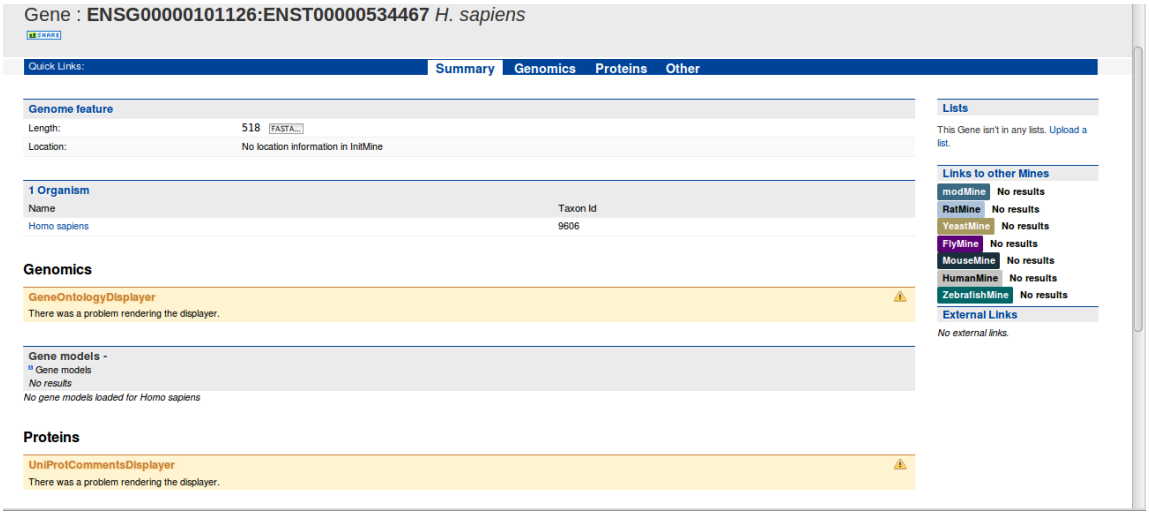


Figure s2: an example page from one of the results, showing the places where various data can be brought into one screen once the database is fully configured.

Table s1: List of components ordered for the construction of a server capable of running Intermine

1 x GIGABYTE GA-78LMT-USB3 AMD 760G (Socket AM3+) Micro-ATX Motherboard : £33.20 excl. VAT
1 x Crucial 8GB Memory Module PC3-12800 1600MHz DDR3 Unbuffered CL11 240-pin DIMM : £25.69 excl. VAT
1 x XFX 430 Watt 80+ Bronze Computer ATX Power Supply : £25.60 excl. VAT
1 x AMD (Piledriver) FX-6300 3.50GHz (4.10GHz Turbo) Socket AM3+ 6-Core Processor - Retail : £65.98 excl. VAT
1 x Arctic Cooling Freezer 7 PRO Rev.2 CPU Air Cooler : £12.04 excl. VAT
1 x Aria E-commerce Shipping Charge : £8.95 excl. VAT
1 x desktop computer case: £2 donation to the SoMakeIt hackerspace

Due to the purchase of other, unrelated items in the same order, and due to time constraints, VAT calculations are omitted.