Pathway prediction by bioinformatic analysis of the untranslated regions of the CFTR mRNA

Jean Spence

Omnitron Biosciences, P.O. Box 601002, San Diego, CA 92160, USA

ARTICLE INFO

Article history:
Received 19 January 2009
Accepted 11 March 2009
Available online 21 March 2009

Keywords:
3′ UTR
5′ UTR
Motif
Bioinformatics
Cystic fibrosis
CFTR

ABSTRACT

Mining the information contained within the genetic code in untranslated regions has proven difficult because of the ambiguity of microRNA and protein binding sites. This manuscript describes a bioinformatic screen that identifies long sequences with partial identity to the untranslated regions of the cystic fibrosis transmembrane regulator. This screen uncovered a long, evolutionarily conserved motif common to the 3′ UTRs of the CFTR and SEC24A transcripts, and shorter, statistically significant motifs unique to either 5′ or 3′ UTRs. In addition, of the 140 transcripts identified in the screen that encode proteins with known protein interactions, 130 are linked to CFTR through protein interactions. The screen identified genes that are known to be involved in lung fibrosis, the inflammatory response of cystic fibrosis and sensitivity to Pseudomonas aeruginosa infections. The bioinformatic analysis of untranslated regions should prove to be a powerful adjunct to other tools for predicting pathways and relevant interactions.

© 2009 Elsevier Inc. All rights reserved.

Introduction

The use of bioinformatics to predict pathways and gene expression is a valuable tool in biology. A potential resource for predicting pathways lies in the bioinformatic analysis of untranslated regions of mRNA. 5′ and 3′ UTRs contain regulatory motifs that determine the location and timing of translation, the rate and efficiency of translation, and the stability of the transcript [1–4]. Additional sequences are involved in splicing and export of the transcripts from the nucleus to the cytoplasm [1,2]. Recently, it has been proposed that transcripts that encode proteins that function in a common pathway are processed as a cohort from transcription through splicing, export from the nucleus, and translation [5]. The coordinating processing of these transcripts is regulated, in part, through the interaction of RNA binding proteins and microRNAs with short nucleotide sequences [5,6].

The identification of the regulatory sequences in untranslated regions has been complicated by the ambiguity of the binding sequences for many micro RNAs and RNA binding proteins. Localization signals that have been identified in 3′ UTRs are also highly variable in sequence and length and may encompass sequences of more than 100 bases [7]. Although microRNA and protein binding sites tend to be relatively short and heterogeneous, long sections of hundreds to thousands of bases in 3′ UTRs are often highly conserved throughout evolution indicating that there are additional regulatory mechanisms [8,9]. In this manuscript, I propose that long sequences with partial identity may be used to identify regulatory sequences in 3′ UTRs in order to identify common pathways or proteins that undergo similar processing during translation. The 3′ UTR for the cystic fibrosis transmembrane conductance regulator (CFTR) was chosen for the initial screen for sequences in the human messenger RNA collection from ReSeq that have partial identity. CFTR is a CAMP-regulated chloride ion channel structurally related to ATP-binding cassette proteins [10]. Cystic fibrosis (CF) is the most common lethal genetic disease in Caucasian populations at approximately 1 in 2500 births [11]. Typically, although many organs are affected, most patients succumb to repeated bouts of inflammation and Pseudomonas aeruginosa infection of the lungs, resulting in death at 40 years [11,12]. The CFTR 3′ UTR is conserved in mammalian genomes and partially in birds and extensive evidence indicates that translational control contributes significantly to the maturation of the protein [13]. With the most common CFTR mutation, ΔF508, nearly 100% of the mutant CFTR protein fails to fold properly and is degraded by the ubiquitin-proteasomal degradation machinery [14,15]. Once correctly folded, CFTRΔF508 is partially functional, indicating that facilitating correct protein folding may ameliorate the disease symptoms [16]. During translation of the CFTR mRNA, the ribosome docks with the translocon, a protein complex that transfers the nascent protein into the endoplasmic reticulum [13]. In addition, the nascent CFTR is associated with a dynamic protein complex and is not released from the ribosome and translocon after the completion of translation until an energy-dependent maturation process requiring nucleotide triphosphates and cytosol is completed [13]. A second critical step in the maturation of the CFTR protein that is defective in CFTRΔF508 is exit from the endoplasmic reticulum [17,18].
Proper folding may be only one of a number of complex events that contribute to the pathology of cystic fibrosis. Even with identical mutations in the CFTR gene, the severity of the disease fluctuates dramatically between individuals indicating that other ‘modulator’ proteins such as Transforming Growth Factor Beta and Mannose Binding Lectin can compensate for the loss of CFTR [19,20]. Therefore, the identities of proteins that interact with or modulate the activity, maturation or stability of CFTR are critical to understanding the disease. This manuscript describes a ‘sliding gate’ bioinformatic screen that identifies long sequences of 40 or more nucleotides (nt) with 70% or greater identity. The sliding gate screen takes each successive, overlapping fragment of the CFTR 3′ or 5′ UTR and slides over the human mRNA cDNA library or the 3′ UTR library of overlapping sequences. In this way, the regions of identity can be distributed throughout the culled sequences, rather than concentrated in shorter, contiguous sections. The use of secondary structure has not been included in the study because it depends on unknown interacting proteins or microRNAs. Subsequent analysis of the sequences recovered in the screen by several bioinformatic platforms uncovered novel motifs in the CFTR 5′ and 3′ UTRs that may regulate the translation of interacting proteins or structurally related proteins. By expanding these motifs, we have identified sequences in the most evolutionarily conserved region of the CFTR 3′ UTR that have partial identity to sequences in the 3′ UTR of SEC24. SEC24 recognizes and physically associates with CFTR during translocation out of the endoplasmic reticulum [17,18]. In addition, CFTR is linked to 130 out of 140 proteins encoded by mRNAs with partial identity to the CFTR 3′ UTR. Many of these proteins function in pathways germane to the pathology of cystic fibrosis including inflammation, diabetes, fibrosis, male fertility, the response to free radicals, oxidative stress and Pseudomonas aeruginosa infections.

Results and discussion

Evolutionary conservation of the CFTR 3′ UTR

Conservation through evolution is an indication of functional importance, as the conserved sequence indicates that mutations are deleterious. The CFTR 3′ UTR is conserved over evolution through mammals, with a core sequence being conserved in birds (Gallus gallus). Representative percent identities are 99.1% for chimpanzee, 75.0% for cat, 71.1% for cow, 64.8% for mouse and 52.1% for opossum. The most highly conserved sequence lies between nucleotides 5645 to 5826 in the Human CFTR mRNA (NM_000492) and is 38.1% identical between the human, dog, rhesus monkey, horse, armadillo, mouse, opossum, platypus and chicken 3′ UTRs. The conserved residues in the mouse and human sequences are shown in Fig. 1C.

Results of the homology search of the RNA databases

For the human RNA database, 603 unique fragments were identified with 70% or greater identity (0.002% of the library) after eliminating CFTR sequences. Correspondingly, 335 unique fragments with 70% identity to CFTR were recovered from the 3′ UTR library (0.003%). The fragments from the 3UTR.Hum library and the RefSeq library were pooled and collated into 389 sequences (See Supplemental Table S1). Sequences with the highest level of identity (78%) were found in the 3′ UTRs of armadillo repeat containing 8 (ARMC8), zinc finger protein 518A (ZNF518A), insulin receptor substrate 1 (IRS1), Rieske Fe–S domain containing (RFESD), ATP-binding cassette, subfamily DALD member 3 (ABCD3), succinate dehydrogenase complex, subunit C (SDHC), synaptotagmin× (SYT10), coiled-coil domain containing 112 (CCDC112) and coiled-coil domain containing 85A (CCDC85A).
(CCDC85A) and the coding sequence of echinoderm microtubule associated protein like (EML4). The sequences from ARMC8, ZNF518A, SDHC, CCDC112, CCDC85A and EML4 were also among the 20 longest collated sequences with 70% or greater identity to CFTR, further indicating that these sequences may be significant. Of the 389 sequences, 76.1% are located in the 3′ UTR of the transcript, 19.5% are in the coding sequences, 1.5% is in the 5′ UTRs, and 0.5% overlap coding and UTRs. Two percent are found in pseudogenes. Subgroups of proteins that are statistically over-represented by two or more standard deviations are the zinc finger proteins, ankyrin repeat domain proteins and the coiled-coil domain containing proteins.

The relative positions of the regions of identity between the CFTR 3′ UTR and the recovered sequences are shown in Fig. 1A. The distribution of fragments is not random (P < 0.001) and there are four clusters with sequences beginning between 100 and 122, between 1100 and 1126, between 1157 and 1200, and between 1494 and 1517 nt from the stop codon. The second, third and fourth regions with the greatest number of hits correspond to the regions of greatest identity between human and mouse sequences (Fig. 1C). The coding sequences appear to be distributed more randomly (P = 0.006, Fig. 1B), but there is some clustering among sequences beginning at 174 to 196 and 1085 to 1140.

The CFTR 5′ UTR was similarly used to search the RefSeq library. In this case, only 19 transcripts were detected when 54 would be expected if both screens were random (Supplemental Table S2). Out of 19 sequences, 2 were in the 5′ UTR, 10 were in coding sequences, and 7 were in 3′ UTRs, reflecting a lack of bias towards 5′ UTR sequences.

### Identification of an abundant 3′ UTR motif associated with inflammation

The first motif for the sequences between nucleotides 125 and 141 from the stop codon is UUUUUAAAAAAGAAA as determined by MEME (Fig. 2A). The 3′ UTRs that have the greatest identity to these regions are from the 3′ UTRs of ARMC8, ABCD2 and SDHC. A BLAST search with this region from the CFTR 3′ UTR identified a 26 nt sequence from the snail homologue 2 (SNAIL2) 3′ UTR that has 24/26 identical residues (Expect = 0.005). This sequence, ‘AA[GA]UU[GU] UUUUUAAAAAAGAAAACAUU’ encompasses and expands upon the core motif. Single nucleotide polymorphisms have been reported for deletions of A and or U at the center of the palindromic UUUUUUAAAAAA (rs63107760, rs4148726; NCBI SNP).

The evolutionary conservation of this region, the palindrome structure and the large number of 3′ UTR sequences with partial identity strongly indicate that this region is a functional motif. This sequence lies in the first 217 nucleotides of the CFTR 3′ UTR which has previously been identified as conferring mRNA instability [21]. A number of the transcripts with identity to this region have identity to other sequences, as determined by BLAST searches. These include sequences from the 3′ UTRs of ABCD3, C14orf153, IL23R, KIFC1, MAPK6, NMI, PABPN1, PRKACA, STK32B, USP7, and ZNF576. Fig. 2B shows an interaction map of a combination of protein interactions and common 3′ UTR motifs that connects SMAD signaling/inflammatory response (IL23R, SMAD2, NMI, GTF2I, SMURF1 and CXCL13), protein chaperoning and ubiquitination (HPH1, USP7 and UBE2K) and ion transport (CACNG4, SLC9A11 and CFTR).

### Motif and combination protein interaction and sequence identity map for transcripts with identity to CFTR 3′ UTR sequences between 94 and 164

The logos for additional motifs are shown in Table 1. With the exception of the motif derived from nt 1450 to 1486, the sequences are highly AU rich. AU rich regions are involved in transcript stability and in regulating the role of microRNAs in translation [6, 22]. Some of the motifs occur in transcripts with published connections to CFTR or

### Table 1

<table>
<thead>
<tr>
<th>CFTR 3′ UTR Sequences</th>
<th>Number of Transcripts</th>
<th>E-value</th>
<th>Motif</th>
</tr>
</thead>
<tbody>
<tr>
<td>123-151</td>
<td>72</td>
<td>2.0e-134</td>
<td></td>
</tr>
</tbody>
</table>

### Fig. 2

Motif and combination protein interaction and sequence identity map for transcripts with identity to CFTR 3′ UTR sequences between 94 and 164. (A) The motif identified by MEME analysis of the CFTR sequence and the sequences that have identity to the CFTR 3′ UTR sequences between 123 and 153. The E-value was calculated by the MEME software. (B) Network map for proteins associated with the CFTR 3′ UTR sequences between 94 and 164 nt from the stop codon. Blue nodes are proteins encoded by transcripts that were identified in the original screen. Yellow nodes were identified by BLAST searches with the sequences from the original screen. Red lines (edges) that are labeled 3′ UTR connect transcripts that have sequence identity according to BLAST. Pink nodes are common interacting proteins from NCBI Entrez Gene that do not share sequence motifs. CFTR is red.
The sequences between nt 1108 and 1152 in the 3′ UTR of CFTR has been identified by using ClustalW alignments scores and N-J trees with branch lengths. Additional sequences were identified by BLAST or pattern matches. The log represents the alignment of the CFTR 5′ or 3′ UTR sequences and sequences in the listed transcripts. Unless indicated, the sequences are in the 3′ UTR. The fold increase and Z scores are given. Positive is in the sense orientation, negative is the complement. a Additional motif candidates were found by screening RefSeq with pattern matches.

Motifs were identified by using ClustalW alignment scores and N-J trees with branch lengths. Additional sequences were identified by BLAST or pattern matches. The log represents the alignment of the CFTR 5′ or 3′ UTR sequences and sequences in the listed transcripts. Unless indicated, the sequences are in the 3′ UTR. The fold increase and Z scores are given. Positive is in the sense orientation, negative is the complement.

a Additional motif candidates were found by screening RefSeq with pattern matches based on the ClustalW alignments.

b PLA2G16 was identified by a BLAST search with the CRIP1 sequences (Expect = 0.037).

c LRRK2 (Expect = 0.006), DHODH (Expect = 0.006) and CCDC112 (Expect = 0.003) were identified in the original sliding gate screen and by a BLAST search with the CFTR sequence.

d PRKA2A, ITGB8 and NT5DC1 were identified by a BLAST search with the CCDC112 sequences (Expect = 0.006 to 0.007).

cystic fibrosis. The second motif is shared with MIERI1, PEL12 and Myosin 5A. While MYOSB interacts with CFTR [23], CFTR also associates with RAB27A [24] which interacts with MYOSA [25]. Another sequence with plausible connections to CFTR is the CREM/SLC9A11/SLC4A4 motif in nt 1091–1115. CREM (cyclic AMP response modulator) is a b-ZIP transcription factor that mediates cAMP-mediated signal transduction. CFTR is regulated by cAMP [23] and has been found to interact with other solute carrier family members including SLC4A7, SLC4A8, SLC9A3R1 and SLC9A3R2 [26–28]. This motif also highlights the potential complexity of the information involved in translation; motifs may be in untranslatable regions or coding sequences or in the positive or negative orientations.

The sequences between nt 1108 and 1152 in the CFTR 3′ UTR have significant identity, both by BLAST and the sliding gate screen, to 3′ UTR sequences in DHODH (dihydroorotate dehydrogenase; E = 0.006), LRRK2 (leucine-rich repeat kinase 2; E = 0.006) and CCDC112 (Coiled-coil domain containing 112; E = 0.003). Mutations in LRRK2 are associated with Parkinson disease-8 [29], so a correlation with cystic fibrosis is not obvious. Indeed, an alignment of the four sequences by ClustalW found that CCDC112 and DHODH were closer to the CFTR sequence (Table 1, motif 1108–1145). The core motif from this sequence (AA[AU]UUUAUUUUAAAAAUNU) is found in 15 additional sequences, including LRRK2 and LRR2 that were identified by the sliding gate screen. Interestingly, CCDC112 was found to have identity to 3′ UTR sequences from AMP kinase alpha 2 (PRKAA2; Expect = 0.006; Identities = 44/56). Considerable literature connects AMP-activated protein kinase with CFTR activity, but PRKAA1, not PRKAA2 is reported to bind to CFTR in airway epithelial cells [30] which leaves the possibility of an interaction of CFTR with AMK2 in other cell types. The CCDC112 sequence is also closely related to sequences in the 3′ UTRs of ITGB8 (integrin beta 8; Expect = 0.007) and NT5DC1 (5′-nucleotidase domain containing 1; Expect = 0.007) (Table 1).

The motif common to CFTR, HLA-DQA1, PKIA and LOC731682 is the most highly conserved of the motifs with homology to the chicken CFTR 3′ UTR. PKIA inhibits the catalytic subunit of cAMP dependent protein kinase A, a regulator of CFTR activity. Two potentially relevant 3′ UTRs have overlapping motifs with identity to the terminus of the 3′ UTR. Insulin receptor 1 (IRS1) has 20/24 or 30/40 identical residues to the CFTR sequences between nt 1499 and 1535. The oxidation resistance 1 (OXR1) 3′ UTR sequence has 27/34 identical residues identical to nt 1512–1546 of the CFTR 3′ UTR, overlapping the IRS1 sequences. Recovery from oxidative damage and susceptibility to cystic fibrosis related diabetes are hallmarks of CF [11]. The possible microRNA binding sites do not overlap the motifs in Table 1 with the exception of the motif at 1450–1486 that overlaps a potential microRNA binding site for has-miR-499a and has-miR-499b.

Transcripts with multiple 3′ UTR regions of identity

Although only 389 unique sequences were identified, several reside in different locations in the same transcript, indicating that these transcripts in particular, may be regulated or processed through the same pathways as the CFTR transcript. The transcripts with three 3′ UTR sequences with 70% or greater identity to the CFTR 3′ UTR are transmembrane protein 106B (TMEM106B), lymphocyte antigen 75 (LY75), map kinase 6 (MAPK6), and polycorn group finger 5 (PCGF5). Because some of the sequences with identities to the same two regions of the CFTR 3′ UTR may be in the same regulatory family, these were screened by ClustalW for common motifs. Two such pairs were found: CALML4 (calmodulin-like 4) and NARG2 (NMDA receptor regulated 2), and MAPK6 and LY75 (Supplemental Fig. S1). Although CALML4 and NARG2 are not well studied, calmodulin regulates the activity NMDA receptors [31]. MAPK6 binds to MAP2 (microtubule associated protein 2; Supplemental Table S1) and mediates glucose stimulated insulin secretion in rat islet cells [32] and thus may contribute to cystic fibrosis related diabetes.

Notably, four transcripts had two regions of identity to the same region of the CFTR 3′ UTR (nt 1187–1239). The sequences from copine III (CNPE3), rap guanine nucleotide exchange factor 4 (RAPGEF4), chromosome 12 open reading frame 40 (C1orf40) and TMEM106B are shown in Supplemental Table S3. All eight of these sequences are from the 3′ UTRs of their respective transcripts. Moreover, a BLAST search identified additional 3′ UTR sequences that were closely related to some of these sequences. The evolutionary conservation of this region between 1187 and 1239 and the recovery of related 3′ UTR sequences by multiple techniques (BLAST, pattern matching and sliding gate string match) indicate that this region contains a functional motif.

Identification of extended evolutionarily conserved extended motifs

In order to determine if the 41 plus nucleotides that have 70% identity are part of a longer motif with regions of lower identity, the 3′ UTR motifs in Supplemental Table S1 were used to extract an additional 50 nucleotides on each side, extending the sequence length to 140 or more nucleotides. These were then screened for long motifs against an alignment of the CFTR 3′ UTRs from a combination of human, rhesus macaque, dog, horse, mouse, armadillo, platypus and...
Identification of a motif in the 5′ UTR of CFTR

In order to determine if the 5′ UTR of CFTR had potential motifs, the 5′ UTR sequence was cut into nested 41 nucleotide sequences and screened against the human mRNA RefSeq library. The sequences that have 70% or greater identity to the CFTR 5′ UTR are shown in the Supplementary Table S2. Of these, only two, CAMKK2 (calcium/calmodulin-dependent protein kinase 2) and BTR20 (zinc finger and BTB domain containing 20), have related sequences in the 5′ UTR of their transcripts. However, none of the sequences at 70% identity were found in transcripts that were identified in the 3′ UTR screen (Supplemental Table, S1). Transcripts that had partial identity of 5′ UTR sequences to the CFTR 5′ UTR at a 62% or greater level of identity (Supplemental Table, S1) included the 5′ UTR and initial coding sequences of cystopolymy binding protein 3 (PABPC3) and the 5′ UTR of coiled-coil domain containing 113 (CCDC113). These sequences have identity to the nucleotides between nucleotides 78 and 130 in the CFTR 5′ UTR and their 3′ UTR sequences have identity to overlapping regions in the CFTR 3′ UTR. Additional sequences with partial identity were found in the coding sequences of TNRC5B, CNTNAP4, DST, MIERI, MY09B, and TLL2 and the 3′ UTR of MOSPD1. The CAMKK2 5′ UTR sequences also overlap the region identified as having identity to PABPC3 and CCDC113. However, the motifs may not be related; the 5′ UTR of CAMKK2 has four repeats of the motif CCGAG (with one mismatch), which is also found in the 5′ UTRs of CHMP4B, GRK6, DD11, ATG9A, C1orf128, and PKR1. The alignment of these sequences resulted in 17 identical matches without the CFTR 5′ UTR (G[AG]CCGACGGAGCCG[AC]GC[AG][AG]) but had only 9 identities with CFTR 5′ UTR, resulting in a Z score below the cutoff point of 10. The CFTR 5′ UTR has the initial copy of the CCGAG repeat and partial identity to the following nucleotides. A pattern search of the RefSeq human mRNA database for sequences using the most compact regions of identity to CCDC113 did not find any related sequences. However, the motif ‘GGNNCGR[AG][G][G][G]CAGNCNN’ was found in the 5′ UTRs of CFTR, platelet-derived growth factor α (PDGFA), and myeloid/lymphoid or mixed-lineage leukemia triatorhomolog 1 (MLLT10) and the 5′ UTR and initial coding sequences of cystopolymy binding protein C3 (PABPC3). The logo for this motif is shown in Table 1. The level of identity and the presence of this sequence solely in 5′ UTRs statistically indicate a function. Mutations or polymorphisms associated with this region include a polymorphism of C to T at position 99 that was found in a patient with disseminated bronchiectasis (www3.genet.sickkids.on.ca/cftr/) and the polymorphism 125G that has been linked to the incidence of idiopathic chronic pancreatitis in Chinese patients [33].

Motifs with identity to coding sequences

Approximately 19.5% of the sequences that were culled in the sliding gate screen are in the coding sequences. Two of the motifs

| Human SEC24A | UAUCAUGCUGUAGUAUAACUUGAGUAUAAAA |
| Human SEC24A | UAAUGCUCCUUUGAUGAAGAAGAAGAAGA |
| Human CFTR | UACUGCUCCUUUGAUGAGGAAGAAGA |
| Mouse CFTR | UAUUGCUCCUUUGAUGAGGAAGAAGA |
| Armadillo CFTR | UAUUGCUCCUUUGAUGAGGAAGAAGA |
| Platypus CFTR | UAUUGCUCCUUUGAUGAGGAAGAAGA |
| Chicken CFTR | UAUUGCUCCUUUGAUGAGGAAGAAGA |
| Human SEC24A | GGCAGAUGCAGAAGAAGAAGAAGAAGA |
| Human SEC24A | GGCAGAUGCAGAAGAAGAAGAAGAAGA |
| Human CFTR | GGCAGAUGCAGAAGAAGAAGAAGAAGA |
| Mouse CFTR | GGCAGAUGCAGAAGAAGAAGAAGAAGA |
| Armadillo CFTR | GGCAGAUGCAGAAGAAGAAGAAGAAGA |
| Platypus CFTR | GGCAGAUGCAGAAGAAGAAGAAGAAGA |
| Chicken CFTR | GGCAGAUGCAGAAGAAGAAGAAGAAGA |
| Human SEC24A | UAUUGCUCCUUUGAUGAGGAAGAAGA |
| Human SEC24A | UAUUGCUCCUUUGAUGAGGAAGAAGA |
| Human CFTR | UAUUGCUCCUUUGAUGAGGAAGAAGA |
| Mouse CFTR | UAUUGCUCCUUUGAUGAGGAAGAAGA |
| Armadillo CFTR | UAUUGCUCCUUUGAUGAGGAAGAAGA |
| Platypus CFTR | UAUUGCUCCUUUGAUGAGGAAGAAGA |
| Chicken CFTR | UAUUGCUCCUUUGAUGAGGAAGAAGA |

Fig. 3. Alignment of the human and mouse SEC24 3′ UTR with the CFTR 3′ UTRs from human (Homo sapiens), rhesus (Macaca mulatta), mouse (Mus musculus), armadillo (Dasyus novemcinctus), platypus (Orinthorhynchus anatinus), and chicken (Gallus gallus). Residues that are identical to the human CFTR 3′ UTR are white on black background. SEC24A residues that are not identical to the CFTR 3′ UTR but that are conserved between the mouse and human SEC24A transcripts are highlighted in gray.
occur either twice in the *CFTR* 3' UTR sequence or 3 times in the coding sequence indicating potential significance. The *CFTR* sequences between 25 and 70 nt and 55–100 nt from the stop codon are present in a tandem overlapping duplication that occurred fairly recently in evolution as it is present in rhesus monkeys but not in the non-primate species analyzed. This region has little identity to other sequences in Refseq but was found to have 71% identity to coding sequences from *EML4* (echinoderm microtubule associated protein like 4). A pattern match based on the identities in the *EML4* and *CFTR* sequences recovered coding sequences in multiple C2 domains, transmembrane 2 (*MCTP2*) (Fig. 4A). The duplication of these sequences in the *CFTR* 3' UTR, the uniqueness of the sequences and the identity in two coding sequences suggests that if functional, these sequences may play a very specific role in coordinating translation, perhaps during embryogenesis.

The second sequence of 49 nt has 69% identity between the *CFTR* 3' UTR and three separate sequences in the coding sequences of *FCGBP* (Fc fragment of IgG binding protein) (Fig. 4B). In the most concentrated region of identity, 19 out of 24 nt are identical in all four sequences (P–E –46). Like the preceding sequence, the sliding gate screen and pattern searches found few additional sequences with identity to this region. *FCGBP* has three tandem in-frame repeats. The coding sequence of *TIPARP* has some overlapping identity to this region (data not shown).

### Protein interactions

In order to investigate the potential relationship between *CFTR* and the proteins encoded by transcripts with sequence identity to the *CFTR* 3' UTR, the protein interactions as defined in NCBI Entrez Gene were collated and mapped. Only 140 out of 389 total transcripts had listed protein interactions. However, only 10 of these could not be tied to the other proteins through one common interacting protein or direct interactions. The map showing over 270 interactions between *CFTR*, the remaining 130 proteins and 78 common interacting proteins is presented in Fig. 5. The map is biased towards proteins that are heavily researched (HIF1A, SMAD2, CHUK/IKKα, etc.). Thirty one of the proteins (24%) interact directly and the remaining proteins share one or more interacting proteins from Entrez Gene. *CFTR* does not interact directly with any of the gene products identified by the bioinformatic analysis but five *CFTR* interacting proteins (EZR, PDZK1, SLC9A3R2, SLC9A3R1 and PRKCE) interact with a subset of proteins encoded by transcripts recovered in the screen. Of the 7 proteins that interact with these proteins, PLCB1 and PRKCA are of interest as phospholipase C and protein kinase A regulate the activity of *CFTR* [23]. SGK3, which interacts with SLC9A3R2, regulates the activities of potassium and sodium channels including ENaC, a cystic fibrosis modulator that functions in concert with *CFTR* to maintain salt concentrations [34,35]. PALLD is mutated in familial pancreatic cancer [36] and linked to *CFTR* through Ezrin and could be involved in the pancreatic deficiency associated with CF.

Notably, of the 11 clusters of screened proteins that interact, six have published connections to *CFTR* and one other is potentially relevant. The largest cluster of interacting proteins, consisting of PTPrC, IRS1, CHUK/IKKα, SMAD2, PPM1B, BIRC3 and RBL1, is involved in inflammation and insulin metabolism. The NFkB pathway, which is activated by IKKα and regulated by SMAD2 and SMAD3, has been shown to be the signaling pathway responsible for inflammation in the lungs of cystic fibrosis patients [12]. SMAD2 and SMAD3 mRNA and protein levels are significantly reduced in *CFTR* −/− mice [37]. SMAD3 also acts downstream from the cystic fibrosis modulator gene, TGF-β, and regulates inflammation and fibrosis [38]. IKKα is over-expressed in CF bronchial cells and has been investigated as a target for Interleukin-10 therapy [39]. Chronic inflammation and recurring infections of the lungs are considered to be the primary contributors to lung damage in cystic fibrosis [12]. Thus, it is intriguing that PIAS1 and UBE2I have been identified as IKKα activates PIAS1 by phosphorylation in a reaction that requires the SUMO ligase, thereby inhibiting the expression of pro-inflammatory genes [40]. PIAS1 and STAT1 are over-expressed and found as a complex in CF cells [41] and a close link between *CFTR* and UBE2I, the SUMO conjugating enzyme exists as both interact with PDZK1 (NCBI Entrez Gene). This could be relevant to the controversy over the origin of the inflammatory response in CF; whether it is due to over expression of pro-inflammatory genes or the failure to turn off inflammation when infections subside [12].

Two other clusters include regulatory proteins that control the activity of *CFTR*. HIF1A (HIF1A/ZNF197) negatively regulates *CFTR* [42] and protein kinase A (PRKACA/PKIA) positively regulates *CFTR* [23]. FGR2A (FGR2A/LGALS3) is a modulator as genetic variations have been associated with increased susceptibility to *P. aeruginosa* [43]. LGALS3, which interacts with FGR2A, has been reported to bind to *P. aeruginosa* [44] and assist in the correct folding of CFTRα508 [45]. Another germane protein is FGF7 (fibroblast growth factor 7 or keratinoctye growth factor) which has been advocated as a gene therapy target for CF based on positive results in vitro [46,47]. Platelet derived growth factor B (PDGFB) and collagen, type 1, α1 (COL1A1), which interact with FGF7, are involved in fibrosis in the lung but have not been published in connection with cystic fibrosis [48,49]. The interacting cluster with potential importance based on the literature is MYO5A/TRIM2. MYO5A interacts with RAB27A, a negative regulator of *CFTR* [24,25]. Other proteins with published interactions are IL23R, the receptor for IL-23 which is critical for the inflammatory response to *P. aeruginosa* in cystic fibrosis [50], the regulatory subunit A of protein phosphatase 2A (PPP2R1B) [23] and GJC1, which physically interacts with *CFTR* at gap junctions [51]. NFAT is over-expressed in CD4+ T cells from *CFTR* −/− mice and may contribute to the inflammatory response or susceptibility to *P. aeruginosa* infections.

![Fig. 4. Alignment of coding sequences with the *CFTR* 3' UTR.](image-url)

(A) Echinoderm microtubule associated protein like 4 (*EML4*) and multiple C2 domains, transmembrane 2 (*MCTP2*) coding sequences are aligned with an overlapping tandem repeat sequence in the *CFTR* 3' UTR. (B) Alignment of three separate repeats from the coding sequence of Fc fragment of IgG binding protein and the *CFTR* 3' UTR.
The high relevance of the selected proteins that directly interact renders the others more compelling.

Conclusion

Use of the deciphered code of UTRs has the potential to be a powerful tool in systems biology for predicting pathways if transcripts that contain these codes are processed coordinately and encode proteins that function in the same process or phase in differentiation. In order to be applicable, a bioinformatics screen should detect both motifs specific to UTRs and transcripts encoding proteins that interact. The most striking 3′ UTR motif resides in the 3′ UTR of SEC24A and CFTR. In addition, the screen has identified several motifs that have low probabilities of chance and high Z scores that were only found in untranslated regions and were recognized by several techniques. Examples include the 5′ UTR motif (Table 1), the motif ‘UUUUUUUAAAAAAGAAA’ (Fig. 2A), and the submotif ‘AA[AU]NUUUAUUUAAAAAUNU’ (Table 1).

The cytoscape map in Fig. 5 illustrates the protein interactions of 130 out of 140 proteins with known interactions in NCBI Entrez Gene. Many of the proteins exhibit multiple interactions with other proteins indicating that the screen has selected a cluster of interrelated pathways. In addition to transcripts that are known to be involved in cystic fibrosis, the screen also uncovered several transcripts that could be relevant. As CFTR is a chloride transporter, other chloride channels may compensate for the lack of CFTR function. OSTM1 is an obligatory subunit of the chloride ion channel-7 and the cause of a severe form of osteopetrosis when mutated [53]. As CFTR is regulated by cAMP and G proteins [23], PKIA, the inhibitor of PRKACA, RAPGEF4, CREM, RGNEF (Rho-guanine nucleotide exchange factor) and GNA11 (guanine nucleotide binding protein G protein, α inhibiting activity polypeptide 1) have potential as regulatory proteins in CF. CREM, in particular, as a male fertility indicator gene, may play a role in the infertility in male patients [54]. While ubiquitin specific proteases USP7, USP28, USP46, and USP48 were selected by the sliding gate screen, a high content proteomic screen recently found that under-expression of USP47 enables mutated CFTR to fold more efficiently [45]. Cystic fibrosis patients often develop cystic fibrosis related diabetes making IRS1 and insulin-like growth factor 2 mRNA binding protein 1 (IGF2BP1) of interest.

Some of the transcripts with the strongest matches encode mitochondrial proteins DHODH and DUT which are essential for pyrimidine biosynthesis and SDHC (succinate dehydrogenase C), a critical component of the tricarboxylic acid cycle and electron transport [55]. Another selected nuclear gene encodes mitochondrial protein ALDH5A1 (NAD(+)-dependent semialdehyde dehydrogenase) which is essential for γ amino butyrate (GABA) synthesis, by converting succinate semialdehyde to succinate [55]. The GABA shunt is coupled to the Krebs cycle and also glutamate and glutamine synthesis [55]. Glutamate in the absence of ATP allows the CFTR to gate Cl− whereas, when bound by ATP, CFTR conducts both Cl− and HCO3− [56]. The failure to conduct HCO3− is associated with the most severe CF cases [56]. Glutamine, the precursor of glutamate, is reduced in the neutrophils of severe CF patients [57] and the cerebrospinal fluid of ALDH5A1 deficient patients [58]. If glutamate and ATP synthesis are altered by certain CFTR mutations, it would result in an overall decrease in metabolism and the effect could be synergistic on CFTR and further reduce ion transport. Signaling between CFTR and
the mitochondrial proteins could be mediated through CDC112 and AMP Kinase Alpha 2 which share a motif with CFTR. Finally, the screen recovered several proteins that function in the immune response in addition to FCGR2A, including major histocompatibility complex, class II, DQ α1 (HLA–DQA1) and Fc fragment of IgG binding protein (FCGBP). Other HLA class II antigens have been identified as possible modifier genes in CF [19].

In conclusion, the sliding gate bioinformatic screen identified statistically significant motifs and an unusually high number of interacting proteins with relevance to CF. The motifs tend to be relatively long and contain gaps, fitting the current understanding of eukaryotic motifs [59]. The screen successfully selected clusters of proteins that are associated with CF, including proteins currently under consideration as therapeutic targets (IKKα, FGF7) and possible modulator proteins (FRG2A, LGALS3). The results also portend an extremely complex interweave of signals involving untranslated regions, coding sequences and motifs in plus and minus orientations. The later would occur if extended duplexes are involved in translational control. The analysis of untranslated regions can be a powerful contribution to the bioinformatics of mapping pathways.

Materials and methods

Databases

Two RNA databases were used: the database of intact mRNA transcripts, human.rna.gbf.gz, was downloaded from ncbi.nih.gov/refseq and the database of non-redundant human 3′ UTRs (3′UTR, Human.rtr.dat.gz) was downloaded from http://bighost.area.ba.cnri.it/BIG/UTRHome/. The nucleic acid sequences were parsed from the databases such that each transcript represented an element in an array. The DNA sequences from human.rna.gbf.gz were obtained by splitting the library into a hash with the identifying information as the key and then recovering the values. The human 3′ UTR library was split into an array with the identifying information and the sequences as separate elements. The sequences were then parsed from the identifying information. The sizes of the databases in nucleotides are 40,143,577 nt for the human 3′ UTR library and 91,526,013 nt for the human RNA library from National Center for Biotechnology Information (NCBI). The transcripts were then processed individually as array elements into 41 nucleotide segments that overlapped by 3 nt. The size of the library of overlapping fragments from the RefSeq human RNA library was 28,273,013 strings and the Human 3′ UTR library consisted of 12,904,525 strings. The 3′ UTR of CFTR was similarly digested into 41 nt sections that overlap by one nucleotide.

Bioinformatic screen

Using a ‘sliding gate’ string matching program written in Perl, the CFTR 3′ UTR library was compared to the genomic transcripts and fragments with 70% identity were returned. Small nucleolar RNAs were analyzed separately with the same string match program. Redundant sequence fragments were eliminated and the names of the genes were recovered. Overlapping fragments were collated and the sequences were mapped according to their respective positions on the CFTR 3′ UTR by using Perl scripts. Their functional location in the transcript (5′ UTR, coding sequence, 3′ UTR) was confirmed by using NCBI BLAST. Additional motifs were found by aligning the 3′ UTRs with ClustalW (http://align.genome.jp/) and using the identical nucleic acids to construct a regular expression for a Perl script. Single base mismatches were allowed for longer motifs. Motifs are written by using A,G,C or U and N to represent any base. Positions that can be occupied by one of two bases encode the two bases in brackets, e.g. [AG] represents A or G.

MEME analysis was accessed online at http://meme.sdsc.edu/meme. Sequence Alignments were performed by ClustalW and Logos were generated by weblogo at http://weblogo.berkeley.edu/. Evolutionary conservation was determined by a BLAT search on the UCSC genome browser (http://genome.ucsc.edu/) which is capable of genome alignment of 28 species. Protein interactions were obtained from NCBI Entrez Gene for each protein, when documented. The protein interaction map was constructed by using Cytoscape which was downloaded from the Cytoscape website. Potential microRNA binding sites were located by the miRBase Target Database (http://microrna.sanger.ac.uk/). Single nucleotide polymorphisms (SNPs) were located in the Cystic Fibrosis Mutation Database at www.genet.sickkids.on.ca/cftr/ and the NCBI SNP database at www.ncbi.nlm.nih.gov/sites/entrez?db=snp.

Statistical analysis

The probability of chance for motif positions in the CFTR 3′ UTR (Fig. 1) was determined by calculating the chi square and using a chi square statistical chart to find the P value and by confirming the results with Excel. The fold increase and Z scores for motifs were calculated according to the manuscript by Xie et al. [60]. In summary, the fold increase was calculated by dividing the incidence of the motif (C) by the expected incidence (μ). Fold enrichment = C/μ. Z score = (C−μ)/sqrt (μ). The expected incidence was calculated according to Xie et al. based on Table 6, Supplemental Information [60]. Because the unmatched nucleotides in the motifs can be any nucleotide in this analysis, these positions were evaluated as having a probability of 1. However, a gap penalty for gaps in alignment was imposed by multiplying the expected incidence by 5 for each gap in alignment. In this case, the results are corroborated by the Expect scores calculated by the BLAST software. For motifs found by using MEME, the P value calculated by MEME is given.

Acknowledgments

The author would like to acknowledge David Rothwarf (Arrayomics) and Carl Ewig (Coleman University) for comments and John Leek for support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jgyeno.2009.03.002.

References

W.B. Guggino, B.A. Stanton, New insights into cystic M.C. Chang, et al., Spectrum of mutations and variants/haplotypes of CFTR and
F. Stanke, et al., The TNFalpha receptor TNFRSF1A and genes encoding the M.C. Frith, N.F. Saunders, B. Kobe, T.L. Bailey, Discovering sequence motifs with
M.M. Reddy, P.M. Quinton, Control of dynamic CFTR selectivity by glutamate and inﬂammatory stimuli induce IKKalpha-mediated phosphorylation of PIAS1 to restrict inﬂammation and immunity, Cell 129 (2007) 903–914.