Published in final edited form as:

Clin Chem Lab Med. 2011 September ; 49(9): 1447–1451. doi:10.1515/CCLM.2011.604.

A template for mutational data analysis of the *CFTR* gene

Giampiero Ferraguti, Silvia Pierandrei, Sabina Maria Bruno, Fabrizio Ceci, Roberto Strom, and Marco Lucarelli*

Department of Cellular Biotechnologies and Hematology, Sapienza University of Rome, Rome, Italy

Abstract

Background—Automated DNA sequencing produces large amounts of data that need to be analyzed by appropriate software. Personalization of software can be a difficult and time-consuming task, especially if a large number of mutations have to be analyzed.

Methods—The Applied BioSystems SeqScape software, based on the KB basecaller algorithm, is a versatile tool that can be used for mutational analysis and for data quality assessment of sequences belonging to any gene of interest. Using this software we analyzed over 1400 sequences of *CFTR* exons and adjacent intronic zones, representing over 500,000 bases.

Results—We present an up to date specific template and a linked set of instructions for automated labeling of all point mutations and polymorphisms of the *CFTR* gene, whose mutations cause cystic fibrosis (the most common genetic disease among Caucasian individuals). We also describe our refined software settings for mutational analysis, in order to keep to a minimum the need of manual validation.

Conclusions—The use of our template greatly simplifies the mutational analysis of the *CFTR* gene, reducing human intervention. In our opinion, it might not only be useful to researchers that already perform *CFTR* mutational analysis by sequencing methods but it should also improve the approach in those laboratories that already use ABI PRISM instrumentation for a limited mutational analysis of the *CFTR* gene. Similar mutational templates can also be used for other disease causing genes, thus improving molecular genetics protocols.

Keywords

automated sequence analysis; *CFTR* mutational analysis; cystic fibrosis; SeqScape software

Introduction

DNA sequencing is a powerful diagnostic/prognostic tool. Direct sequencing of a PCR product, known as re-sequencing, is a common practice in genetics and molecular biology

© 2011 by Walter de Gruyter

*Corresponding author: Marco Lucarelli, Sez. Biochimica Clinica, Dip. Biotechnologie Cellulari ed Ematologia c/o Azienda Policlinico Umberto I, Sapienza Università di Roma, Viale Regina Elena 324 00161, Rome, Italy, Phone: +39-06-49970458, Fax: +39-06-4464698, marco.lucarelli@uniroma1.it.

Conflict of interest statement

Authors' conflict of interest disclosure: The authors stated that there are no conflicts of interest regarding the publication of this article. Research support played no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the report for publication.

Research funding: None declared.

Employment or leadership: None declared.

Honorarium: None declared.

worldwide. Such automated methods, especially if performed in 96- or 384-well format, generate huge data sets that need to be analyzed. The use of enzymatic dideoxy chain termination is still the most used sequencing technique; when coupled to the dye terminator chemistry, sequencing reactions can be done in a single tube and run on a gel or in a capillary electrophoresis system (1). Electropherograms and text sequences are obtained, that correspond to single-stranded DNA fragments of expected length. The increased power of modern computers together with the development of sophisticated base calling software, such as phred [the first one endowed with a powerful base-specific quality score (2)] and BaseFinder (3), have made it possible to overcome, in order to produce highly accurate sequences, a major problem in DNA sequencing, namely the correct assignment of conflicting readings. More recently, a new revised base calling software, the KB basecaller, has been released by Applied BioSystem. Comparative statistics on several metrics, such as accuracy-based read lengths, predicted read lengths or clear ranges demonstrate that this KB basecaller provides a substantial increase in read length over phred and reduces the rate of basecalling errors (4). The analysis and validation of very large sets of sequencing data cannot be performed in a relatively short time without the support of automated procedures, namely some adequately instructed data analysis software, in order to, for example, correctly tag any varied base among hundreds of bases. Large genes with multiple sets of sequence variations could, on the other hand, create a tedious, time-consuming charge that offsets the advantage of automated re-sequencing. A typical example of the difficulties that can be encountered in sequence data analysis occurs in molecular diagnostics of cystic fibrosis (CF), a recessive disease (5, 6), where the characterization of the mutations on both alleles can require extensive sequencing of the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene (7) that has a genomic sequence of approximately 250 kb, 27 exons and over 1500 known mutations and polymorphisms (8, 9). In a previous work (10), we proposed a procedure for extensive re-sequencing of the *CFTR* gene; the high throughput nature of such a procedure can be further increased by automation of the sequencing data analysis. In this paper we present a template and the linked set of instructions for *CFTR* re-sequencing data analysis, based on SeqScape (11), one of the most common commercially available (Applied BioSystems) softwares for sequence analysis.

Materials and methods

The last demonstration version of the SeqScape Software (Applied BioSystems) is available at the following address: <http://www.appliedbiosystems.com/support/software>. With the aim of performing a mutational search of the proximal 5' flanking region of *CFTR*, of all exons of this gene and of the adjacent intronic zones, we designed a template that is available on the journal website as supplementary material linked to this article. This is provided as a compressed file that can be downloaded, decompressed as multiple subfiles and imported in the subfolders of SeqScape manager tab according to the set of instructions enclosed in the appropriate paragraph below. We used the SeqScape software version 2.0 and this template can also be imported in the newest 2.7 version. Once all the files have been loaded, experimental sequence data can be analyzed according to the instructions described (the names of the specific functions or sections of SeqScape software are indicated). A SeqScape project can be assembled by importing experimental data into the appropriate exon-specific project template. A project template contains: (a) Reference data groups (RDGs), each RDG analyzes a single exon and the adjacent intronic zones. The reference nucleotide sequence imported from GenBank is organized in intron-exon-intron structure defining a region of interest (ROI) and analytic layers; all the point mutations, sequence variations and amino acidic variants occurring within the selected region are described. (b) Analysis default settings: they specify how the project, specimen and sample analysis settings are applied to all the samples and specimens as they are imported into a project. In the project and specimen subfolder, default settings can be used and in the sample subfolder, a previously

created analysis protocol must be selected. (c) Display settings: they control how projects are visually displayed as soon as the analysis is complete. When samples are loaded into a project, the “auto add” function can be useful. In order to use this function, it is necessary to standardize the sample name by including some delimiters. These can be “X_” as start and “_exon-number” as end (e.g., “X_samplename_exonnumber”). Specimens, containing single or multiple sample runs, after choosing the appropriate layer, are analyzed according to the project template parameters. Reliability of base calling can be quantified in terms of its quality score (2), ranging from a maximum value of 50 down to zero. All bases with a quality value (QV) over 20 can be validated automatically, whereas manual validation can be evaluated for bases with QV below 20 (visualized in yellow or in red if higher or lower than 15, respectively). In our experience with this protocol of extensive re-sequencing (10), over 99% of the bases have a QV over 20. The validated experimental sequence is then directly matched, by the software, with the reference sequence previously exported from GenBank and loaded into the template. The SeqScape software highlights, in a graphical differentiated manner, any discrepancy of the experimental from the “wild type” sequence, representing it either as a mutation or as a sequence variation depending on their classification in the Cystic Fibrosis Mutation Database (8). The final assessment of the results is facilitated by a graphical or a tabular display that is also available for printing. For the analysis of all point mutations and sequence variations of the *CFTR* gene, we generated and tested 55 files organized in three subfolders and grouped in a single .rar archive. These 55 files, that constitute the whole template, are available as supplementary material on the website linked to this article. WinRAR software, to decompress .rar archives, can be downloaded from: <http://www.winrar.it/prelievo.php>. These files are organized in three folders. One folder contains 26 project template files, another 26 reference data groups files and the last includes three files (display settings, analysis protocol, analysis default). The group of 26 project templates files: one file allows analysis of the 5′ flanking zone and of exon 1, another is dedicated to exons 6a and 6b and to the short 6a intron in between whereas the other 24 files can be used for the analysis of single exons. Once decompressed in the appropriate folder (named Project Templates) this group of files can be uploaded by sequentially clicking, in the upper left portion of the main window of the SeqScape software, the following items: Tools and SeqScape Manager. After having opened the SeqScape Manager general tab, choose the Project Templates subfolder and use the import button to find and select files. The group of 26 RDG files: these files are organized similarly to the project templates files and can be uploaded by a procedure similar to that described above in the SeqScape RDG subfolder. The last group is made of display settings, analysis protocol and analysis default files: these three files can also be uploaded into the appropriate SeqScape Manager tabs as mentioned above.

Results

This *CFTR* template for SeqScape is partitioned into the three groups of files described below and can easily be personalized according to specific requirements. The first group contains the 26 project template files that can be used for the analysis of all *CFTR* exons, of the intron-exon boundaries and of the 5′ flanking zone. The second group contains the 26 RDG files, spanning all 27 exons (and the adjacent intronic zones) and the proximal 5′ flanking zone of the *CFTR* gene. The corresponding ROI for each RDG file is organized into layers. Layer one is a blocked, non-modifiable layer where the reference sequence exported from GenBank was loaded. From reporting the GenBank file name, the related accession number and source file can be retrieved by using the NCBI Entrez Nucleotide tool. The published sequence was found to contain some errors as previously reported (10) in the 5′ flanking zone and in introns 8, 17a and 20; in addition errors in intron 3 (405+46T correct, instead of G incorrect) and in intron 22 (4268+112, +CC added) were also found. These errors were confirmed and corrected on the basis of the experimental sequence

obtained for each zone in at least 300 independent samples. In all these cases the correct sequences were loaded on the *CFTR* template (with the corresponding files marked as _corrected). In layers two to five, the reference sequence shows the exonic and the surrounding intronic zones. Layer two, which spans the complete length of the loaded reference sequence, is a general purpose layer that can be used to analyze the sequencing data obtained by any set of primers. By contrast, layers three to five, which span a shorter reference sequence zone, are particularly suitable for the analysis of data obtained by the specific set of primers used in our previously described procedure for the extensive re-sequencing of the *CFTR* gene (10). In particular, these layers allow the best sequencing data analysis that can be obtained starting from amplicons produced by external primers and sequenced by using internal primers. They can be easily modified to analyze sequence data obtained by different sets of external and internal primers including the use of an identical pair of primers for both PCR amplification and sequencing. Layer three is well suited to the analysis of double strand sequence data, whereas layers four and five are well suited for the analysis of forward and reverse single strand sequence data, respectively. Only one layer, selected according to the type of experimental data is usually used for the analysis. If the layer is longer than the experimental sequence, false mutation annotations can appear at either end of the sequence. In this case, the use of the shorter layers between three and five, as opposed to the complete layer two, has the advantage of yielding a more accurate annotation of the mutations in reports and in the graphical interface. Additional layers (from six to 11) were also added to files devoted to analysis of multiple adjacent zones (either the 5' flanking region plus exon 1, or exon 6a, intron 6a and exon 6b) and single zones could thus be analyzed (the 5' flanking region separately from exon 1 and exon 6a separately from exon 6b, respectively). In layers from two to 11, the translation frame and the index codon number were also loaded in order to allow the exact translation in an amino acid (AA) view and the correct classification of AA variants that can be present in the protein. Except for two full intronic mutations, all known point mutations and polymorphisms were loaded into the nucleotide variants (NT variants) section of the RDG file. The corresponding AA variants were also loaded into the appropriate section of the RDG file, but only in layer three (whereas the NT variants can be visualized on all the layers). Insertions and deletions can also be easily detected by the *CFTR* template-based SeqScape analysis but their automated classification proved to be unfeasible. In our *CFTR* template, every molecular alteration was loaded in such a way as to be readily classified either as mutation or as sequence variation with no apparent pathological relevance [and to be therefore reported according to the database of the Consortium for CF Genetic Analysis (8) as "mutation" or "sequence variation" in both the NT and AA variants description]. This preliminary classification can be expressed in printed reports or in a graphical interface (where mutations will appear with a red foreground over yellow background, whereas sequence variations will be denoted with a black foreground over a cyan background) and it can be easily modified on the basis of novel findings. The template can also accommodate novel (or not previously loaded) mutations and sequence variations, classified accordingly to their pathological relevance. However, because the distinction between mutations and sequence variations is often a highly controversial point, the preliminary classification provided by the template should not be uncritically used for diagnostic purposes or for functional relevance assessment. A possible evolution of this template could indeed be its extension by including additional information (based on selected literature reports and/or on experimental results) regarding the pathogenic role, in typical and atypical forms of CF, of every variant listed in the template. The vast majority of mutations and sequence variations classified to date and included in the database of the Consortium for CF Genetic Analysis (8) could be inserted in the template described in this paper (Table 1). Apart from the two full intronic mutations that were not examined, only the large deletions could not be accommodated because they escape any standard PCR based search. The third group of files present in the template is composed of three files containing personalized display settings, analysis protocol and

analysis default. Novelties were introduced in the analysis protocol file. The DyeSet primer and the basecaller used are KB_3100_POP6_BDTv1_SR (susceptible of modification if a different sequencing chemistry is used) and KB, respectively. When dealing with heterozygous samples, we chose for IUPAC (International Union of Pure and Applied Chemistry) mixed base calling of the smaller peak, a threshold set at 25% of the highest peak with both peaks being evaluated by their fluorescence emission and expressed as relative fluorescence units. This setting was obtained by considering the lowest percentage of overlap between peaks retrieved in our experimental DNA sequences, that is, at a somewhat lower value than the lowest percent overlap (30%) experimentally observed for actual mixed base callings in 57 different missense mutations occurring in the *CFTR* gene and found by sequencing both the forward and the reverse strand. This threshold setting was validated by testing sequences with the highest background signals; in our experience, even in these circumstances no incorrect assignment was ever observed. Clear range setting, as determined by QV score, was chosen with the aim of removing bases from each end of DNA sequence until less than one base out of 20 has a QV below 15; with such an automated setting, only few bases (approx. 20) need to be manually corrected at the beginning and at the end of a DNA sequence. A stricter setting would unnecessarily reduce the length of the original sequence. Quality values section of display settings can be changed. In over 500,000 bases analyzed, the nucleotides marked with a red QV bar (that indicates a QV score below 15, corresponding to an accuracy lower than 96%) in the consensus sequence, deduced from two forward and two backward readings, were only 0.78% with only 0.025% of them with a QV score ranging from 6 to 5 (no base ever had a QV <5). These red-targeted bases <4000 in total, escaped the need of any manual correction even when their quality score was 5, which was the lowest quality score found in our experience. Based on these facts, the use of the yellow quality bar (QV below 20 but higher than 15) was practically superfluous, because yellow marked bases could escape manual revision without any problem. Color and ranges of the QV bar can be modified (although not fully eliminated) in the display settings file by entering the SeqScape manager tab. The red bar upper threshold can thus be advantageously shifted to a lower value (e.g., 6 instead of 14) with no risk of any unnoticed erroneous base call, but our suggestion, especially to beginners, is to leave default QV settings prudentially unmodified.

Discussion

Over 1400 sequences were analyzed and validated both manually and in an automated manner by using the SeqScape software with this template and its refined software settings. No errors were generated by the automated procedure as compared to the manual one. More than an additional 2000 sequences were also analyzed and validated by using only the automated procedure, without any apparent mistake. At least 30 *CFTR* gene tracts have to be amplified and individually re-sequenced in order to span all the 27 exons (with the exon 13 being too large to be amplified in a single PCR run and consequently split in two amplicons), the adjacent intronic zones and the 5' flanking zone (amplified in two PCR runs) (10). The analysis of these unavoidable large amounts of *CFTR* re-sequencing data is greatly simplified and becomes dramatically faster by using our template and the linked set of instructions. By downloading these template files, there is no need to perform the tedious task of personalizing SeqScape software for *CFTR* re-sequencing analysis. Our template can be useful to the community of laboratories that perform *CFTR* mutational analysis by gene sequencing methods in typical and atypical forms of this disease. It could also enable laboratories that are already using ABI PRISM instrumentation for a limited mutational analysis of the *CFTR* gene to extend their investigation to the complete analysis of all exons and to their adjacent intronic regions. From a brief survey of the literature and of the existing laboratory facilities in the USA and European countries, at least two-thirds of the laboratories with activities in CF molecular diagnostics usually restrict their investigation to

the most common mutations, failing therefore to detect at least 20% of the *CFTR* mutated alleles. In effect, the first level methods (both in house and commercial) that search for the most frequent *CFTR* mutations generally have a low detection rate (12) and thus a high number of cases have to be analyzed by second level approaches. Our template can be contributory not only if the re-sequencing is the second level method of choice but also if other scanning procedures like denaturing high-performance liquid chromatography (DHPLC), denaturing gradient gel electrophoresis (DGGE) or single-stranded conformation polymorphism (SSCP) are preferred. Also in this case, the *CFTR* re-sequencing of the selected tracts (e.g., those containing the heteroduplexes) has to be performed to identify the mutations. It should also be taken into account that the number of cases approaching the second level tests is growing, due to the ever increasing technological power and the progressive reduction of costs, as well as the number of subjects involved in potential *CFTR* related disorders. According to the updated European recommendations (13), these cases of *CFTR* related disorders have to be often studied with scanning procedures, as first level analyses usually show a detection rate even lower than that shown for classic CF. Moreover, the possible occurrence of complex alleles where the presence of more than a single mutation can be detected only by extended sequencing of the gene, has been found to have some relevance in clarifying phenotype variability (14). Our model of a gene-specific template with the related set of instructions for automated labeling of mutations and polymorphisms, is a general procedure that would be advisable to apply, with specific changes, to the study of other genes. In a more general view, the approach of sharing mutational templates for each disease gene could greatly enhance the mutational search ability of applied molecular genetics protocols.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by grants awarded by “Eleonora Lorillard Spencer Cenci” and Telethon (project GGP06199) Foundations.

References

1. Strausberg RL, Levy S, Rogers YH. Emerging DNA sequencing technologies for human genomic medicine. *Drug Discov Today*. 2008; 13:569–77. [PubMed: 18598911]
2. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*. 1998; 8:175–85. [PubMed: 9521921]
3. Giddings MC, Severin J, Westphall M, Wu J, Smith LM. A software system for data analysis in automated DNA sequencing. *Genome Res*. 1998; 8:644–65. [PubMed: 9647639]
4. Gehman, C.; Labrenz, J.; Baybayan, P.; Cook, LL.; Doan, Q.; Egan, A., et al. [Accessed: 2 September 2010] Longer reads with the KB Basecaller. 2004. Available at: http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_040250.pdf
5. Davies JC, Alton EW, Bush A. Cystic fibrosis. *Br Med J*. 2007; 335:1255–9. [PubMed: 18079549]
6. Ratjen FA. Cystic fibrosis: pathogenesis and future treatment strategies. *Respir Care*. 2009; 54:595–605. [PubMed: 19393104]
7. Riordan JR. CFTR function and prospects for therapy. *Annu Rev Biochem*. 2008; 77:701–26. [PubMed: 18304008]
8. Consortium for CF Genetic Analysis. [Accessed: 16 May 2010] Available at: <http://www.genet.sickkids.on.ca/cftr/app>

9. Bobadilla JL, Macek M Jr, Fine JP, Farrell PM. Cystic fibrosis: a worldwide analysis of CFTR mutations – correlation with incidence data and application to screening. *Hum Mutat.* 2002; 19:575–606. [PubMed: 12007216]
10. Lucarelli M, Narzi L, Piergentili R, Ferraguti G, Grandoni F, Quattrucci S, et al. A 96-well formatted method for exon and exon/intron boundary full sequencing of the CFTR gene. *Anal Biochem.* 2006; 353:226–35. [PubMed: 16635477]
11. Le H, Hinchcliffe M, Yu B, Trent RJ. Computer-assisted reading of DNA sequences. *Methods Mol Med.* 2008; 141:177–97. [PubMed: 18453090]
12. Tomaiuolo R, Spina M, Castaldo G. Molecular diagnosis of cystic fibrosis: comparison of four analytical procedures. *Clin Chem Lab Med.* 2003; 41:26–32. [PubMed: 12636046]
13. Dequeker E, Stuhmann M, Morris MA, Casals T, Castellani C, Claustres M, et al. Best practice guidelines for molecular genetic diagnosis of cystic fibrosis and CFTR-related disorders – updated European recommendations. *Eur J Hum Genet.* 2009; 17:51–65. [PubMed: 18685558]
14. Lucarelli M, Narzi L, Pierandrei S, Bruno SM, Stamato A, d’Avanzo M, et al. A new complex allele of the CFTR gene partially explains the variable phenotype of the L997 mutation. *Genet Med.* 2010; 12:548–55. [PubMed: 20706124]

Table 1

CFTR point mutations, sequence variations and small insertions or deletions suitable for analysis by our template as compared to those classified to date and included in the database of the Consortium for CF Genetic Analysis (8).

Nature of DNA sequence alteration	Total number of sequence alterations present in the database of the Consortium for CF Genetic Analysis	Total number of sequence alterations that can be successfully analyzed by our template
Pathologically relevant point mutations	958	958
Pathologically relevant small insertions or deletions	290	290
Full intronic mutations	2	
Large deletions	47	
Pathologically irrelevant point sequence variations	212	212
Pathologically irrelevant small insertions or deletions	15	15