

OPEN ACCESS Full open access to this and thousands of other papers at http://www.la-press.com.

SHORT REPORT

Blasted Cell Line Names

Jing Wang¹, Lauren A. Byers², John S. Yordy³, Wenbin Liu¹, Li Shen¹, Keith A. Baggerly¹, Uma Giri⁴, Jeffrey N. Myers⁵, K. Kian Ang³, Michael D. Story⁶, Luc Girard⁷, John D. Minna⁸, John V. Heymach² and Kevin R. Coombes¹

¹Department of Bioinformatics and Computational Biology, University of Texas M.D. Anderson Cancer Center, Houston, TX, USA. ²Department of Thoracic/Head and Neck Medical Oncology, University of Texas M.D. Anderson Cancer Center, Houston, TX, USA. ³Department of Radiation Oncology, University of Texas M.D. Anderson Cancer Center, Houston, TX, USA. ⁴Department of Experimental Radiation Oncology, University of Texas M.D. Anderson Cancer Center, Houston, TX, USA. ⁵Department of Head and Neck Surgery, University of Texas M.D. Anderson Cancer Center, Houston, TX, USA. ⁵Department of Head and Neck Surgery, University of Texas M.D. Anderson Cancer Center, Houston, TX, USA. ⁶Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX, USA. ⁷Department of Pharmacology, University of Texas Southwestern Medical Center, Dallas, TX, USA. ⁸Department of Internal Medicine, University of Texas Southwestern Medical Center, Bilas, TX, USA. ⁸Department of Internal Medicine, University of Texas Southwestern Medical Center, Bilas, TX, USA. ⁸Department of Internal Medicine, University of Texas Southwestern Medical Center, Bilas, TX, USA. ⁸Department of Internal Medicine, University of Texas Southwestern Medical Center, Bilas, TX, USA. ⁸Department of Internal Medicine, University of Texas Southwestern Medical Center, Bilas, TX, USA. ⁸Department of Internal Medicine, University of Texas Southwestern Medical Center, Bilas, TX, USA. ⁸Department of Internal Medicine, University of Texas Southwestern Medical Center, Bilas, TX, USA. ⁸Department of Internal Medicine, University of Texas Southwestern Medical Center, Bilas, TX, USA. ⁸Department of Internal Medicine, University of Texas Southwestern Medical Center, Bilas, TX, USA.

Abstract

Background: While trying to integrate multiple data sets collected by different researchers, we noticed that the sample names were frequently entered inconsistently. Most of the variations appeared to involve punctuation, white space, or their absence, at the juncture between alphabetic and numeric portions of the cell line name.

Results: Reasoning that the variant names could be described in terms of mutations or deletions of character strings, we implemented a simple version of the Needleman-Wunsch global sequence alignment algorithm and applied it to the cell line names. All correct matches were found by this procedure. Incorrect matches only occured when a cell line was present in one data set but not in the other. The raw match scores tended to be substantially worse for the incorrect matches.

Conclusions: A simple application of the Needleman-Wunsch global sequence alignment algorithm provides a useful first pass at matching sample names from different data sets.

Keywords: Blast, Cancer cell lines

Cancer Informatics 2010:9 251-255

doi: 10.4137/CIN.S5613

This article is available from http://www.la-press.com.

© the author(s), publisher and licensee Libertas Academica Ltd.

This is an open access article. Unrestricted non-commercial use is permitted provided the original work is properly cited.



While trying to integrate multiple data sets collected by different researchers, we have noticed that the sample names were frequently entered inconsistently. These inconsistencies make it difficult to automate the process of matching data correctly, since matching procedures tend to be based on exact matches of character strings. These inconsistencies can cause problems in a variety of contexts. For example, searching on cell line names at the web site for the American Type Culture Collection (ATCC) can fail if the name copied from a publication is not exactly the same as the name stored in their database.

While the problem of inconsistent names appears to be especially prevalent when using cell lines, our experience also indicates that similar problems can arise when using other kinds of samples as well. For instance, laboratory technicians often add extra information (in the form of a prefix or suffix) to the character string naming the sample in order to annotate something special that happened during sample preparation. In this case, the sample identifiers no longer exactly match the corresponding identifiers in a database containing clinical information about the samples. In an ideal world, of course, all of the extra information would be stored in a database that carefully regulated the forms of identifiers that could be used. In practice, data is often transferred to statisticians or bioinformaticians in spreadsheets or other files that do not adhere to strict standards or naming conventions.

Most of the variations in sample identifiers appear to involve punctuation, white space, or their absence at the juncture between alphabetic and numeric portions of the cell line name. The second most common variations appear to involve suffixes or prefixes added to the\standard" version of the sample identifier. Because the variant names can be described in terms of mutations, deletions, or insertions of character strings, we reasoned that algorithms that had already been developed for alignment of DNA or protein sequences could be applied to the problem of matching cell line names. The most commonly used sequence alignment algorithm at present is the Basic Local Alignment Search Tool (BLAST).¹ However, our task at present is to align fairly short sequences as completely

as possible, and thus a global alignment algorithm seems more appropriate. As a result, we chose to implement a simple version of the original Needleman-Wunsch global sequence alignment algorithm.² Scripts to implement and apply the algorithm in the R statistical software environment³ are available from the authors upon request.

We tried to integrate three kinds of data. The first data set contained radiation response data (obtained by estimating SF2, the surviving fraction after treatment with 2 Gray) on 33 head and neck squamous cell carcinoma (HNSCC) cell lines, 33 lung cancer cell lines, and 63 cell lines related to the NCI60 (also known as the NCI-60). The second data set contained reverse phase protein lysate array (RPPA) data on 224 HNSCC or lung cancer cell lines. The third data set contained Illumina gene expression data on 105 HNSCC or lung cancer cell lines. Complete lists of the cell line names in the three data sets are contained in the supplementary Excel file (namesMatched.xlsx).

We applied the Needleman-Wunsch algorithm to the cell line names, using a mismatch penalty of 2, a gap penalty of 1, and a match score of 2. Since the primary goal of the intended study was to relate gene or protein expression to outcome in the form of radiation response, we applied the algorithm as follows. First, for each cell line name used in the SF2 data set, we computed the Needleman-Wunsch score for all of the cell line names in both the RPPA data set and the Illumina data set. We then recorded all cell lines with maximum score as potential matches. Representative results are shown in Table 1; a complete list of the results is contained in the supplementary Excel file (namesMatched. xlsx). Because the NCI60 cell lines (other than lung cancer lines) were present only in the SF2 data set, they provide useful information about the behavior of the algorithm when no correct matches exist.

We summarize the results in Table 2. For each raw score, we show the number of correct matches and the number of cell line names that could not be matched because there was no valid counterpart in the other data set. The raw match scores tended to be substantially worse for the incorrect (because impossible) matches than for the correct matches. There were no correct matches with a score less

UMSCC17A HNSCC Correct 16 UMSCC17A	
OSC19 LN1 HNSCC Correct 15 OSC19 LN1	
HCC4017 Lung Correct 14 HCC4017	
UMSCC2 HNSCC Correct 12 UMSCC2	
PCI-15A HNSCC Correct 11 PCI15A	
H2009 Lung Correct 10 H2009	
PCI-13 HNSCC Correct 9 PCI13	
HCC1171 Lung Correct 8 H1171	
HN5 HNSCC Correct 6 HN5	
HCC-2998 NCI60 No match 5 HCC2279; HCC2935	
SNB19 NCI60 No match 4 SN1	
HCT116 NCI60 No match 3 HCC4011	
NCI-H23 NCI60 No match 2 H23; PCI-22B	
TK6 NCI60 No match 1 T406	
T47D NCI60 No match 0 H847; T406; TUN7	
SF-268 NCI60 No match –1 S38; SN2	
OVCAR5 NCI60 No match –2 A549; OSC19LN5	
SK-MEL5 NCI60 No match –3 KA-0; KA-3; KA-G; KH-0; KH-3; KH-G; I	KT53; OSC19LN5
IGROV-1 NCI60 No match –4 LKR13; OSC19; PCI13; TR146	
LOX-IVMI NCI60 No match –7 DBL; DLY; LBL; LLY	

Table 1A. Best matches of SF2 cell line names in RPPA data set.

Notes: For a selected set of cell line names in the SF2 radiation response data set, we indicate which subset (HNSCC, Lung, or NCI60) it comes from, whether the method gives a correct result or whether no match is possible, along with the raw Needleman-Wunsch score, and the name(s) achieving the highest score.

Table 1B. Best matches of SF2 cell line names in Illumina data set.

	Set	Result	Score	Best matches				
OSC19 LN2	HNSCC	Correct	18	OSC19 LN2				
UMSCC10B	HNSCC	Correct	15	UMSCC 10B				
HCC1833	Lung	Correct	14	HCC1833				
HCC827	Lung	Correct	12	HCC827				
UMSCC14A	HNSCC	No match	11	UMSCC 14B; UMSCC 11A				
H1299	Lung	Correct	10	H1299				
UMSCC1	HNSCC	No match	9	UMSCC 10B; UMSCC 14B; UMSCC 11A				
A549	NCI60	Correct	8	A549				
H226	Lung	No match	7	H2126				
H23	Lung	Correct	6	H23				
SCC61	HNSCC	No match	5	SQCCY1; HCC461				
H522	Lung	No match	4	H322				
ML-1	NCI60	No match	3	DM-14				
UACC-257	NCI60	No match	2	HCC827				
PC-9	Lung	No match	1	C39				
PCI-15A	HNSCC	No match	0	HCC15				
OVCAR3	NCI60	No match	-1	C39				
786-0	NCI60	No match	-2	TU686; LN686; H1781; HCC78				
NCI-ADR	NCI60	No match	-3	FADU				
IGROV-1	NCI60	No match	-4	DM-14; OSC19				
SKOV-3	NCI60	No match	-5	C39; H23				
EKVX.L	Lung	No match	-6	HBEC30KT; HBEC34KT				
LOX-IVMI	NCI60	No match	-9	DM-14; OSC19; LN686				

Notes: For a selected set of cell line names in the SF2 radiation response data set, we indicate the best matches in the Illumina gene expression data set.



Table	2.	Summary	of	matching	results
-------	----	---------	----	----------	---------

Score	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4
N correct	0	0	0	0	0	0	0	0	0	0	0	1	1	2
N unmatched	1	0	1	2	3	6	5	9	20	22	16	11	9	18
Score	5	6	7	8	9	10	11	12	13	14	15	16	17	18
N correct	2	8	0	24	2	26	6	11	0	6	8	8	0	1
N unmatched	11	3	4	2	2	0	3	0	0	4	0	0	0	0

Notes: For each raw match score, we list the number of cell line names matched correctly from the SF2 data set to the RPPA or Illumina data sets, as well as the number of incorrect matches that resulted because the cell line was not present in the other data sets.

than 2. With a score less than or equal to 5, there were only 6 (5.7%) correct matches compared to 134 (87.0%) names that were impossible to match.

All correct matches were found by this procedure, with only one name providing an ambiguous match. The cell line name "NCI-H23" in the SF2 data set was correctly matched with the name "H23" in the RPPA data set, but incorrectly matched with the name "PCI-22B" in the same data set. Both matches yielded a raw score of 2. Not surprisingly, this suggests that shorter cell line names are more difficult to match correctly and unambiguously. With that single exception, all other incorrect matches only occured when a cell line was present in one data set but not in the other.

The results also suggest that it is impossible to set a cutoff on the score that will ensure that putative matches with at least that score will be correct. For example, the SF2 data set contains cell lines names "OSC19 LN1" and "OSC19 LN2". Only the second of these cell lines is contained in the Illumina data set.

Thus, the best match to the "OSC19 LN1" cell line in the SF2 data set is the cell line "OSC19 LN2", which has a raw match score of 14. Because many cell lines have names that differ only in a single digit, we expect that highly similar but incorrect matches will be common. Note that our current implementation does not correct for differences of case (eg, "CALU1" vs. "Calu-1"). Difference in case can be handled either by forcing everything to upper case or by adding a more elaborate mismatch penalty matrix that imposes smaller penalties for changes in case. Consequently, we view the use of the Needleman-Wunsch sequence alignment algorithm as a first step in the process of correctly matching sample identifiers, especially across data sets containing hundreds of samples. With this tool, it is possible to quickly assemble a spreadsheet that shows the best putative matches (such as the one provided as a supplementary file). Having this file makes it easy for a researcher to scan through and indicate which matches are correct and which are incorrect. If that information is entered as another column in the same spreadsheet, then the resulting documentation can be used directly by statistical software packages to automate the next step in the process of merging the data.

Acknowledgements

This work was supported in part by the Department of Defense grant W81XWH 07 1 0306 02, and by National Institutes of Health/National Cancer Institute grants P50 CA070907, P50 CA097007, and P01 CA006294.

Disclosure

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers of this paper report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.



References

- 1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403–10.
- Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol.* 1970;48(3):443–53.
- 3. R Development Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing 2006.

Publish with Libertas Academica and every scientist working in your field can read your article

"I would like to say that this is the most author-friendly editing process I have experienced in over 150 publications. Thank you most sincerely."

"The communication between your staff and me has been terrific. Whenever progress is made with the manuscript, I receive notice. Quite honestly, I've never had such complete communication with a journal."

"LA is different, and hopefully represents a kind of scientific publication machinery that removes the hurdles from free flow of scientific thought."

Your paper will be:

- Available to your entire community free of charge
- Fairly and quickly peer reviewed
- Yours! You retain copyright

http://www.la-press.com