

Comparison of NMF and semi supervised NMF for detection of cisplatin-induced dinucleotide substitutions

We first provide some background on the relationship between the variant of NMF presented in (Alexandrov et al. 2013a, 2013b) (termed LA-NMF here) and its relationship to semisupervised NMF (ssNMF). Both LA-NMF and ssNMF are basically unsupervised methods that do not provide statistical measures of uncertainty such as confidence intervals or p values. Recall the customary notation for NMF, $V \approx WH$, in which V is the matrix of observed mutational spectra, W is the matrix of mutational signatures, and H is the matrix of "exposures". The ssNMF implementation is a small modification of the LA-NMF code, and differs from LA-NMF only in that it treats W , the signature matrix, as composed of two segments: W_f , which specifies the known, fixed signatures, and W_u , which is computed by NMF. ssNMF then updates only W_u and H . In the context of the current paper, W_f is the experimental cisplatin dinucleotide substitution (DNS) signature.

With ssNMF one can ask the question: "To what extent can the mutations in an observed set of tumors be explained by the action of a specific **known** mutational signature combined with a reasonably small number of additional, unknown signatures?" In the specific context of the current paper, the question is: "To what extent can the DNS spectra of sets of hepatocellular carcinomas (HCCs) and esophageal carcinomas (ESADs) be explained by the action of the experimental cisplatin DNS signature combined with a reasonably small number of additional, unknown signatures?" LA-NMF, on the other hand, is designed to discover previously unknown signatures.

The benefits of the ssNMF approach compared to LA-NMF stem from (1) the fact that signatures discovered by LA-NMF can be heavily influenced by the specific input tumors when it is applied to a relatively small number of tumors, which is obvious in Section 4.3 of Alexandrov (2014) and which is the situation in the current study, and (2) ssNMF assess exactly the signature being tested (in the case the experimental cisplatin DNS signature), rather than a re-estimated approximation. Both LA-NMF and ssNMF require selection of the number of signatures to be discovered, which depends on a tradeoff between estimated reconstruction error and signature stability, as defined in Step 6 of Alexandrov et al. (2013b). In both LA-NMF and ssNMF reconstruction error and signature stability are estimated across the bootstrapping iterations Steps 2, 3, and 4 of Alexandrov et al. (2013b). In both methods the final choice of the number of signatures is left to human judgement. Increasing the number of signatures usually reduces reconstruction error but decreases the stability of the signatures. For ssNMF, if W_f has actually contributed to the observed spectra, then for a given number of signatures, both ssNMF and LA-NMF will have similar estimated reconstruction errors. This is the case in the current study, in which, for 3 signatures total, we have the following, where V_{HCC} and V_{ESAD} are as defined in the main text.

	V_{HCC}		V_{ESAD}	
	ssNMF	LA-NMF	ssNMF	LA-NMF
Signature stability	0.830	0.962	0.842	0.867
Average reconstruction error	329	527	162	177

Indeed, in this case the cisplatin signatures discovered by LA-NMF are very similar to the experimental signatures, with cosine similarities of 0.940 and 0.997 for V_{HCC} and V_{ESAD} , respectively.

In summary, ssNMF is designed to determine the extent to which mutations in a set of tumors can be explained by the action of a specific **known** mutational signature, and in this case its results are confirmed by analysis with LA-NMF.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL et al. 2013a. Signatures of mutational processes in human cancer. *Nature* 500: 415-421.

Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. 2013b. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* 3: 246-259.

Alexandrov LB. 2014. Signatures of mutational processes in human cancer. Thesis submitted for the degree of Doctor of Philosophy, Darwin College, University of Cambridge.