HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2011 July 01.

Published in final edited form as:

Nat Methods. 2011 January ; 8(1): 70–73. doi:10.1038/nmeth.1541.

SAINT: Probabilistic Scoring of Affinity Purification - Mass Spectrometry Data

Hyungwon Choi[1], **Brett Larsen**[2], **Zhen-Yuan Lin**[2], **Ashton Breitkreutz**[2], **Dattatreya Mellacheruvu**[1], **Damian Fermin**[1], **Zhaohui S. Qin**[3], **Mike Tyers**[2,4,5], **Anne-Claude Gingras**[2,4,*], and **Alexey I. Nesvizhskii**[1,6,*]

[1] Department of Pathology, University of Michigan, Ann Arbor, MI 48109-0602, USA

[2] Centre for Systems Biology, Samuel Lunenfeld Research Institute, 600 University Avenue, Toronto, Ontario, M5G 1X5, Canada

[3] Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA

[4] Department of Molecular Genetics, University of Toronto, 1 Kings College Circle, Toronto, Ontario, M5S 1A8, Canada

[5] Wellcome Trust Centre for Cell Biology and Centre for Systems Biology, School of Biological Sciences, University of Edinburgh, Mayfield Road, Edinburgh, EH9 3JR, Scotland, UK

[6] Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109-0602, USA

Abstract

We present SAINT (Significance Analysis of INTeractome), a computational tool that assigns confidence scores to protein-protein interaction data generated using affinity-purification coupled to mass spectrometry (AP-MS). The method utilizes label-free quantitative data and constructs separate distributions for true and false interactions to derive the probability of a *bona fide* protein-protein interaction. We demonstrate that SAINT is applicable to data of different scales and protein connectivity and allows for the transparent analysis of AP-MS data.

The analysis of protein complexes and protein interaction networks is of central importance in biological research. A combination of affinity purification and mass spectrometry (AP-MS) has been increasingly used for both small scale and large scale analysis of protein complexes and interaction networks [1–4]. However, the development of computational tools

*To whom all correspondence should be addressed. gingras@lunenfeld.ca, nesvi@med.umich.edu.
Present address: Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30329, USA

Author contributions

H.C. and A.I.N. developed, implemented and tested the SAINT method; H.C. wrote the software; B.L., A.B., Z.L., A.-C.G. and M.T. generated data for the initial SAINT modeling and provided feedback on the model performance; D.M. and D.F. assisted with data analysis and processing; Z.S.Q. contributed to statistical model development; H.C., A.-C.G. and A.I.N. wrote the manuscript; A.I.N. and A.-C.G. conceived the study; A.I.N directed the project with input from A.-C.G.

SAINT is available as Supplementary Software and at http://saint-apms.sourceforge.net/.

for the processing of AP-MS data has not kept pace with improvement in experimental approaches. In addition to the general challenge of false positive protein identifications in MS-based proteomic data[5], unfiltered AP-MS datasets contain a large number of non specifically binding proteins; filtering these contaminants represents the foremost computational challenge.

While early methods filtered the noise using binary data (presence or absence of a protein), more recently proposed methods take into account quantitative information embedded in the mass spectrometric data (e.g. label-free quantification, such as spectral counts). For example, one recently described method converts the normalized spectral abundance factor (NSAF) into the posterior probability of a true interaction between a bait-prey pair using simple heuristics, which we term PP-NSAF hereafter [6]. Another method, CompPASS computes scores that adjust observed spectral counts relative to the reproducibility of detection across biological replicates and to the frequency of observing prey proteins in purifications of different baits[7]. Although both approaches are effective in analyzing the datasets for which they were developed, these scores are an empirical transformation of spectral counts without a probability model for the measurement errors in the data in a transparent manner.

In a recent work we introduced an advanced approach for statistical analysis of interaction data from AP-MS experiments utilizing label-free quantification, which we termed Significance Analysis of INTeractome (SAINT) [8]. Like PP-NSAF and CompPASS, our original SAINT approach was designed for the analysis of a specific dataset, the yeast kinase and phosphatase interactome. Here we present a generalized SAINT framework that can compute interaction probabilities in a variety of datasets. The method incorporates negative controls commonly generated as a part of the experimental study, but can also be applied to large datasets in the absence of such data. Here we illustrate the methodology and its advantages through the analysis of datasets of different sizes and network density levels: from a large, sparsely connected network involving human deubiquitinating enzymes to a smaller, highly interconnected network for chromatin remodeling proteins, and even to the analysis of a single bait, the protein CDC23.

The aim of SAINT is to convert the label free quantification (spectral count X_{ij}) for a prey protein *i* identified in a purification of bait *j* into the probability of true interaction between the two proteins, $P(True|X_{ij})$. The spectral counts for each prey-bait pair are modeled with a mixture distribution of two components representing true and false interactions. Note that these distributions are specific to each bait-prey pair. The parameters for true and false distributions, $P(X_{ij}|True)$ and $P(X_{ij}|False)$, and the prior probability π_T of true interactions in the dataset, are inferred from the spectral counts for *all* interactions involving prey *i* and bait *j*. SAINT normalizes spectral counts to the length of the proteins and to the total number of spectra in the purification.

In addition to the experimental data for bait proteins, AP-MS data often contain negative controls (Fig. 1a). When these are available, SAINT estimates the spectral count distribution for false interactions directly from the negative controls, which makes the modeling approach semi-supervised (see **Methods**). SAINT modeling can also be performed without

negative control data, given that a sufficient number of independent baits are profiled, and provided that these baits are not densely interconnected. In this case (illustrated in Fig. 1b), a prey detected in the purification of a bait is scored in reference to the quantitative information for the same prey across purifications of all other baits in the dataset. While this is possible for large datasets such as the yeast kinase and phosphatase network[8] and the human deubiquitinating (DUB) enzyme interaction network[7] (that each contain >75 baits; see below), this unsupervised approach involves additional assumptions and separate treatment of high and low frequency prey proteins (see **Methods**).

One challenge in modeling AP-MS data is the limited number of replicates available for each bait. SAINT addresses this problem by inferring individual bait-prey interaction parameters via joint modeling of the entire bait-prey data. To this end, SAINT defines a protein-specific abundance parameter and establishes a multiplicative model in the mixture component distributions. In other words, if prey *i* and bait *j* interact, then the "interaction abundance" (the spectral count of the prey *i* in purification with bait *j*) is assumed to be proportional to $a_i \times a_j$. Under this assumption, the protein-specific abundance parameters a_i and a_j can be learned not only from the interaction between the two proteins themselves, but also from other *bona fide* interactions that involve either one of them. The same principle applies to false interactions. Hence SAINT builds a large number of mixture distributions by pooling data (separate mixture distributions for individual prey-bait pairs), but all models are interconnected through the shared abundance parameters.

The probability distributions $P(X_{ij}|True)$ and $P(X_{ij}|False)$ are then used to calculate the posterior probability of true interaction $P(True|X_{ij})$ (Fig. 1c and 1d, **Methods**). For baits profiled in replicates, the next step involves computing a combined probability score from independent scoring of each replicate (see **Methods**). Finally, SAINT probabilities can be used to estimate the false discovery rate (FDR). By ordering interactions in a decreasing order of probabilities, a threshold can be selected that considers the average of the complement probabilities as the Bayesian FDR[9]. Although the accuracy of FDR estimates remains to be validated, the availability of an objective reliability measure that has been widely used is an advantage over other methods.

The performance of the generalized SAINT model was first investigated using a human dataset centered around four key protein complexes involved in chromatin remodeling, Prefoldin, hINO80, SRCAP, and TRRAP/TIP60 (referred to as the TIP49 dataset)[6]. While the original publication focused the analysis on the interaction network observed between a core set of 65 proteins, the entire dataset provided by the authors of the study is analyzed here. The dataset consists of 27 baits (35 purifications) and 1207 preys which yielded 5521 unfiltered interactions. 35 negative controls were included in the dataset, allowing semi-supervised modeling (Fig. 1a; Supplementary Table 1).

We applied SAINT to this data and compared the results to PP-NSAF[6] and CompPASS Z and D^N scores[7,10], which we re-implemented in-house (see **Methods**). We note that PP-NSAF[6] removes all interactions involving prey proteins for which the sum of squared NSAF values across the negative control purifications is higher than that in the experiments

containing bait proteins. CompPASS is the only method that does not incorporate negative controls in scoring.

SAINT selected 1375 interactions at the probability threshold 0.9, which was approximately equivalent to an estimated FDR of 2%. In PP-NSAF, since arbitrary cutoffs were set to define high, moderate, and low probability interaction sets, the same number of top scoring interactions was selected from the method (corresponding to a PP-NSAF probability 0.2 or higher). In CompPASS, the same number of interactions corresponded to a D^N-score threshold of 1.48 (Supplementary Table 1).

We evaluated the performance of each algorithm firstly by benchmarking the selected interactions against two interaction databases BioGRID[11] and iRefWeb[12] (Fig. 2a), and secondly by assessing the co-annotation rate of interaction partners to common Gene Ontology (GO) terms in Biological Processes (Fig. 2b; Supplementary Table 1). SAINT filtered interactions (with controls) consistently showed the highest overlap with previously reported interactions and co-annotation rates to terms relevant to chromatin remodelling, including histone acetylation, protein amino acid acetylation, chromatin organization and modification, and cellular macromolecular complex assembly. Variation of the SAINT probability thresholds (0.8 ~ 0.95) did not qualitatively change this conclusion (data not shown). Note that omission of negative controls from SAINT modeling decreased the literature overlap (Supplementary Fig. 1). Explicit incorporation of the negative control data improves the robustness of modeling, especially in small to medium datasets.

The performance of SAINT for large scale datasets without negative controls (Fig. 1b) was tested on the human deubiquitinating enzymes (DUB) dataset [7] (this dataset was used in the development of CompPASS). High confidence interactions from SAINT were compared to the high confidence set from CompPASS (see Supplementary Table 2). Due to the absence of negative controls, it was not possible to apply PP-NSAF to this dataset. SAINT probabilities and D^N scores were notably correlated (Pearson correlation r=0.79). At probability 0.8 threshold, SAINT selected 1300 interactions, while CompPASS $D^N \geq 1$ (threshold value used in [7]) reported 1377 interactions. Of these, 1051 interactions were common to both methods. Reflecting the similarity of selected interactions, SAINT and CompPASS recovered previously reported interactions at comparable rates (Fig. 2c). In the top 1000 interactions, SAINT showed higher overlap with literature data. The co-annotation of interaction partners to the common GO terms also showed similar results between the two methods (Fig. 2d), including relevant terms such as positive and negative regulation of ubiquitin-protein ligase activity during mitotic cell cycle, proteasome, etc. (Supplementary Table 2). While SAINT and CompPASS recovered largely overlapping interactions, SAINT removed the interactions identified with 1–2 spectral counts, which were still scored by CompPASS if they were specific to a single bait protein and detected in duplicates.

Another advantage of SAINT over other methods is that it is applicable to the analysis of small-scale datasets for which control purifications are available; this extends to the case of a single bait. We illustrate this by using a recent dataset[13] containing 3 experimental purifications of the bait CDC23 and 3 control purifications. In the original analysis, the authors of the study identified true interactions using ion intensity-based quantification

followed by a simple t-test. We applied the SAINT approach to the same dataset by using spectral counts (the data was researched in-house as described in **Methods**). The results obtained by SAINT were nearly identical to the initial report (Supplementary Table 3), the sole exception being the single peptide hit C11orf51, which was reported as a new interactor in the original analysis[13], but which was removed by SAINT.

In summary, SAINT is a probability-based model that is generally applicable to mass spectrometry-based interaction data. The SAINT model presented here is based on label-free quantification using spectral counts, a parameter that is easily extracted from most AP-MS datasets. However, SAINT can also be extended to model other types of quantitative parameters such as peptide ion intensity [14] or other continuous variables [15], which can be accommodated by simply substituting the likelihood with an appropriate continuous distribution.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Supported by grants from the CIHR to A.C.G. (MOP-84314), and M.T. (MOP-12246), the NIH to M.T. (5R01RR024031), to A.I.N. and A.-C.G. (R01-GM094231), and A.I.N. (R01-CA126239), a Royal Society Wolfson Research Merit Award and a Scottish Universities Life Sciences Alliance Research Chair to M.T, a Canada Research Chair in Functional Proteomics to A.C.G., and the Lea Reichmann Chair in Cancer Proteomics to A.C.G. The authors thank Mihaela Sardiu, Mike Washburn, and Matthew Sowa for providing additional information regarding the datasets used in this work, Gary Bader for helpful discussions and Wade Dunham for critical reading of the manuscript.

References

1. Ewing RM, et al. Mol Syst Biol. 2007; 3
2. Gavin AC, et al. Nature. 2006; 440:631–636. [PubMed: 16429126]
3. Jeronimo C, et al. Mol Cell. 2007; 27:262–274. [PubMed: 17643375]
4. Krogan NJ, et al. Nature. 2006; 440:637–643. [PubMed: 16554755]
5. Nesvizhskii AI, Vitek O, Aebersold R. Nat Methods. 2007; 4:787–797. [PubMed: 17901868]
6. Sardiu ME, et al. Proc Natl Acad Sci U S A. 2008; 105:1454–1459. [PubMed: 18218781]
7. Sowa ME, Bennett EJ, Gygi SP, Harper JW. Cell. 2009; 138:389–403. [PubMed: 19615732]
8. Breitkreutz A, et al. Science. 2010; 328:1043–1046. [PubMed: 20489023]
9. Muller, P.; Parmigiani, G.; Rice, K. Bayesian Statistics. Bernardo, JM., et al., editors. Vol. 8. Oxford Univ. Press; 2007.
10. Behrends C, Sowa ME, Gygi SP, Harper JW. Nature. 2010; 466:68–76. [PubMed: 20562859]
11. Breitkreutz BJ, et al. Nucleic Acids Res. 2008; 36:D637–640. [PubMed: 18000002]
12. Turner B, et al. Database (Oxford). 2010; 2010:baq023. [PubMed: 20940177]
13. Hubner NC, et al. J Cell Biol. 2010; 189:739–754. [PubMed: 20479470]
14. Rinner O, et al. Nat Biotechnol. 2007; 25:345–352. [PubMed: 17322870]
15. Griffin NM, et al. Nat Biotechnol. 2010; 28:83–89. [PubMed: 20010810]
16. Eng JK, McCormack AL, Yates JRI. J Am Soc Mass Spectrom. 1994; 5:976–989. [PubMed: 24226387]
17. Ishwaran H, James LF. J Am Stat Assoc. 2001; 96:161–173.

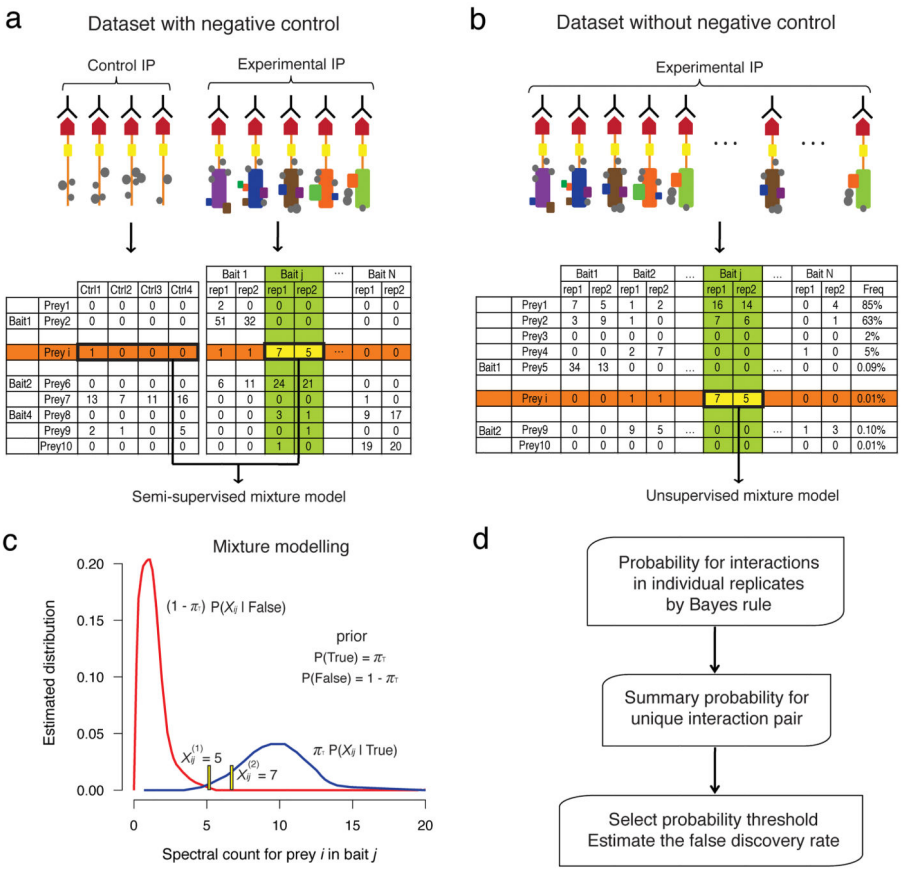

Figure 1. Probability model in SAINT

a–b Interaction data in the presence (**a**) and absence (**b**) of control purifications. *Top*: schematic of the experimental AP-MS procedure; *Bottom*: illustration of a spectral count interaction table. **c**. Modeling spectral count distributions for true and false interactions. For the interaction between prey *i* and bait *j*, SAINT utilizes all relevant data for the two proteins, as shown in the column of the bait (green) and the data in the row of the prey (orange) in **a** and **b. d.** Probability is calculated for each replicate by application of Bayes rule, and a summary probability is calculated for the interaction pair (*i*,*j*).

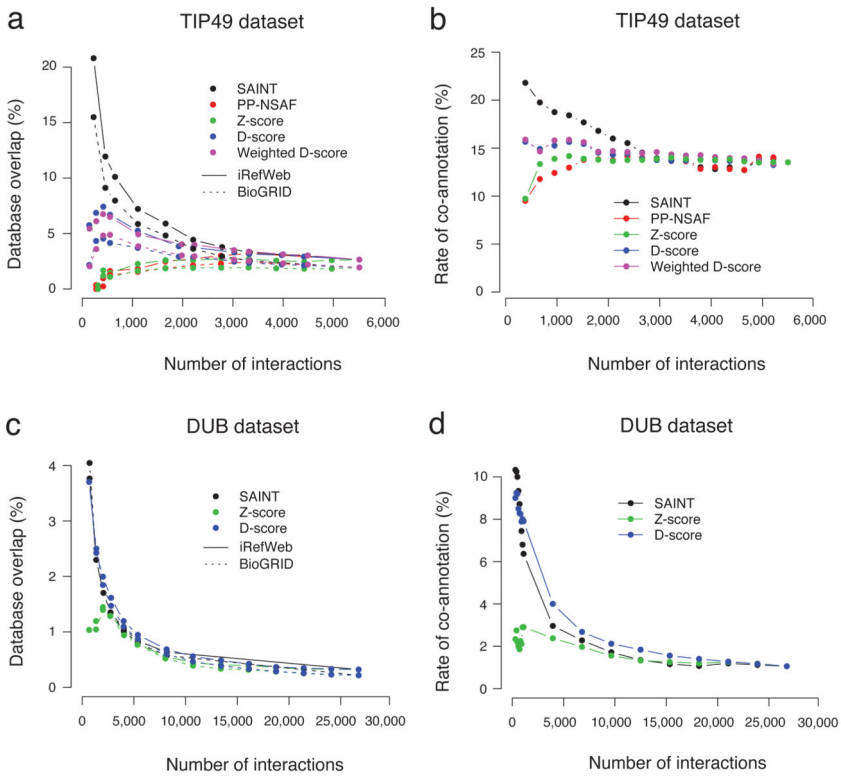

Figure 2. Analysis of TIP49 and DUB datasets

a. Benchmarking of filtered interactions in the TIP49 dataset by the overlap with interactions previously reported in BioGRID and iRefWeb databases. **b**. Co-annotation of interaction partners to common GO terms in Biological Processes in the TIP49 dataset. **c**. Benchmarking against BioGRID and iRefWeb in the DUB dataset. **d**. Co-annotation to GO terms in the DUB dataset.