Supervised Nonparametric Image Parcellation

Mert R. Sabuncu[1], B.T. Thomas Yeo[1], Koen Van Leemput[1,2,3],
Bruce Fischl[1,2], and Polina Golland[1]

[1] Computer Science and Artificial Intelligence Lab, MIT
[2] Department of Radiology, Harvard Medical School
[3] Dept. of Information and Computer Science, Helsinki University of Technology

Abstract. Segmentation of medical images is commonly formulated as a supervised learning problem, where manually labeled training data are summarized using a parametric atlas. Summarizing the data alleviates the computational burden at the expense of possibly losing valuable information on inter-subject variability. This paper presents a novel framework for Supervised Nonparametric Image Parcellation (SNIP). SNIP models the intensity and label images as samples of a joint distribution estimated from the training data in a non-parametric fashion. By capitalizing on recently developed fast and robust pairwise image alignment tools, SNIP employs the *entire* training data to segment a new image via Expectation Maximization. The use of multiple registrations increases robustness to occasional registration failures. We report experiments on 39 volumetric brain MRI scans with manual labels for the white matter, cortex and subcortical structures. SNIP yields better segmentation than state-of-the-art algorithms in multiple regions of interest.

1 Introduction

Image segmentation in medical imaging aims to partition images into various regions of interest (ROIs), such as anatomical structures. Except in cases where the ROIs are distinguishable based on intensity information alone, prior information is typically needed in the form of manually labeled data. A common approach is to summarize the training data with a parametric model, usually referred to as an *atlas* [1,2,3,4,5]. Atlases aid segmentation by introducing a global coordinate system that restricts the number of possible structures occurring at a particular position and may encode the appearance of anatomical structures.

Atlas-based segmentation relies on the alignment of a new image to the atlas coordinate frame. Conventional methods utilize off-the-shelf inter-subject registration tools as pre-processing before segmentation [6,7,3]. Because the quality of registration can be improved with better segmentation and vice versa, several approaches have been proposed to unify the two problems [8,4,5].

An alternative strategy is to employ the entire training data set. Such an approach can exploit recently-developed fast and accurate, pairwise nonlinear registration algorithms, e.g. [9,10]. The label fusion (propagation) method [11,12] transfers the labels of training images to a test image after pairwise registration.

G.-Z. Yang et al. (Eds.): MICCAI 2009, Part II, LNCS 5762, pp. 1075–1083, 2009.

The segmentation labels of the test image are then estimated via majority voting. This method yields improved segmentation, since errors in the registration procedures are averaged out. A recent extension of label fusion [11] uses a subset of the training data, consisting of the subjects most similar to the test subject. Yet, segmentation is still performed via majority voting, where each relevant training subject has the same weight. Isgum *et al.* propose an ad-hoc method that uses local and soft weighting within the label-fusion framework [13].

In this paper, we develop a supervised nonparametric image parcellation (SNIP) framework conceptually similar to label fusion [12] and its extensions [11,13]. In contrast to these methods, we adopt a Bayesian approach, where segmentation is inferred via the Maximum A Posteriori (MAP) principle and the joint label and intensity image distribution is estimated in a nonparametric fashion. The transformations between the test image and each training image are modeled as nuisance random variables and marginalized using standard Bayesian approximations. Marginalization accounts for the uncertainty in registration, commonly ignored in the literature (see [14,15] for notable exceptions). The resulting optimization is efficiently solved using Expectation Maximization. Unlike [12], the similarity between a warped training image and test image plays an important role: more similar training images are weighted more in segmentation.

The soft weighting of training subjects was recently used for shape regression [16], where the weights were a function of age difference between the subjects. The proposed SNIP framework is also related to STAPLE [17], which fuses multiple segmentations of a single subject. In contrast, SNIP handles multiple subjects and accounts for inter-subject registration.

We report experiments on 39 brain MRI scans that have corresponding manual labels, including the cortex, white matter, and sub-cortical structures. We demonstrate that SNIP compares favorably to state-of-the-art segmentation algorithms in multiple regions of interest.

2 Theory

Let $\{I_i\}$ be N training images with corresponding label maps $\{L_i\}$, $i = 1, \ldots, N$. We assume the label maps take discrete values that indicate the label identity at each spatial location. Let $I : \Omega \mapsto \mathbb{R}$ denote a new, previously unseen test image defined on a discrete grid $\Omega \subset \mathbb{R}^3$. One common approach to estimate its label map $\hat{L}$ is via MAP estimation:

$$\hat{L} = \underset{L}{\operatorname{argmax}}\, p(L|I, \{L_i, I_i\}) = \underset{L}{\operatorname{argmax}}\, p(L, I|\{L_i, I_i\}), \tag{1}$$

where $p(L, I|\{L_i, I_i\})$ denotes the joint probability of the label map L and image I given the training data. Rather than using a parametric model for $p(L, I|\{L_i, I_i\})$, we employ a non-parametric estimate:

$$p(L, I|\{L_i, I_i\}) = \frac{1}{N} \sum_{i=1}^{N} p(L, I|L_i, I_i). \tag{2}$$

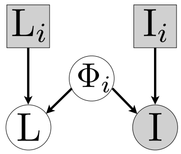

Fig. 1. Generative model for (L, I) given the template (L_i, I_i). Φ_i is the mapping from the image coordinates to the template coordinates. Squares indicate non-random parameters, while circles indicate random variables. Shaded variables are observed.

Eq. (2) can be viewed as a kernel density (Parzen window) estimate of the underlying distribution, or equivalently, a mixture distribution. $p(L, I|L_i, I_i)$ is the probability of (L, I), given that the new subject was generated from training subject i. Let $\Phi_i : \Omega \mapsto \mathbb{R}^3$ denote the unknown transformation that maps a test image grid point to a location in the training image i. Fig. 1 illustrates the generative model for $p(L, I|L_i, I_i)$, which assumes that the image I and label L are conditionally independent if the transformation Φ_i is observed. This yields:

$$\begin{aligned} p(L, I|L_i, I_i) &= p(I|L_i, I_i)p(L|I, L_i, I_i) \\ &= p(I|I_i) \int_{\Phi_i} p(L|\Phi_i, L_i, I_i)p(\Phi_i|I, L_i, I_i)d\Phi_i \\ &\approx p(I|I_i) \int_{\Phi_i} p(L|\Phi_i, L_i, I_i)\delta(\Phi_i - \Phi_i^*)d\Phi_i \\ &= p(I|I_i)p(L|\Phi_i^*, L_i), \end{aligned} \tag{3}$$

where we used the standard mode approximation for the integral and

$$\Phi_i^* \triangleq \underset{\Phi}{\operatorname{argmax}}\, p(\Phi_i|I, L_i, I_i) = \underset{\Phi}{\operatorname{argmax}}\, p(I|\Phi, I_i)p(\Phi), \tag{4}$$

is the most likely transformation between test image I and training image I_i. Substituting Eqs. (2) and (3) into Eq. (1) yields

$$\hat{L} = \underset{L}{\operatorname{argmax}} \sum_{i=1}^{N} p(I|I_i)p(L|\Phi_i^*, L_i). \tag{5}$$

The objective function in Eq. (5) can be viewed as a mixture distribution, where the label likelihood terms $p(L|\Phi_i^*, L_i)$ are the mixture components and the image likelihood terms $p(I|I_i)$ – which encode the similarity between the test image I and training image I_i – are the mixing coefficients. This optimization problem can be solved efficiently using Expectation Maximization (EM). In the next section, we instantiate the model and present the corresponding EM algorithm.

3 Model Instantiation

In our current implementation, we assume the image likelihood is a spatially independent Gaussian with a stationary variance σ^2:

$$p(I|I_i, \Phi_i) = \prod_{x \in \Omega} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(I\left(x\right) - I_i\left(\Phi_i\left(x\right)\right)\right)^2\right]. \tag{6}$$

We model the label likelihoods as a product of independent multinomials:

$$p(L|L_i, \Phi_i) = \prod_{x \in \Omega} \pi_{L_i}(L(x); \Phi_i(x)), \tag{7}$$

where $\pi_{L_i}(l; \Phi_i(x))$ encodes the probability of observing label l at grid location $x \in \Omega$ of the test image, given that the test image is generated by training image i and Φ_i is the mapping from the coordinates of the image to those of the training image i. We compute $\pi_{L_i}(\cdot; \Phi_i(x))$ by applying the transformation Φ_i to the vector image $\pi_{L_i}(\cdot; x)$ where each voxel is assigned a length-$\mathcal{L}$ probability vector, with one indicating the manual label, and zero elsewhere. Non-grid values are obtained via trilinear interpolation.

Using the one-parameter subgroup of diffeomorphism, we parameterize a warp Φ with a smooth, stationary velocity field $v : \mathbb{R}^3 \mapsto \mathbb{R}^3$ via an ODE [9]: $\frac{\partial \Phi(x,t)}{\partial t} = v(\Phi(x,t))$ and initial condition $\Phi(x, 0) = x$. The deformation $\Phi(x) = \exp(v)(x)$ can be computed efficiently using scaling and squaring and inverted by using the negative of the velocity field: $\Phi^{-1} = \exp(-v)$ [18].

We impose an elastic-like regularization on the stationary velocity field:

$$p(\Phi = \exp(v)) = \frac{1}{Z_\lambda} \exp\left[-\lambda \sum_{y \in \Omega} \sum_{j,k=1,2,3} \left(\frac{\partial^2}{\partial x_j^2} v_k(x)\bigg|_{x=y}\right)^2\right], \tag{8}$$

where $\lambda > 0$ is the warp stiffness parameter, Z_λ is a partition function that depends only on λ, and sub-scripts denote coordinates (dimensions). A higher warp stiffness parameter yields more rigid warps.

3.1 Efficient Pairwise Registration

To evaluate the joint probability in Eq. (3), we need to compute Φ_i^* defined in Eq. (4). Using Eqs. (6) and (8), we can rewrite Eq. (4) as

$$\hat{v}^i = \underset{v}{\operatorname{argmin}} \sum_{y \in \Omega} \left[(I(y) - I_i(\exp(v)(y)))^2 + 2\lambda\sigma^2 \sum_{j,k=1,2,3} \left(\frac{\partial^2}{\partial x_j^2} v_k(x)\bigg|_{x=y}\right)^2\right], \tag{9}$$

where $\Phi_i^* = \exp(\hat{v}^i)$. To solve Eq. (9), we use the bidirectional log-domain Demons framework [10], which decouples the optimization of the first and second terms by introducing an auxiliary transformation. The update warp is first computed using the Gauss-Newton method. The regularization is achieved by smoothing the updated warp parameters. The smoothing kernel corresponding to Eq. (8) can be approximated with a Gaussian: $K(x) \propto \exp(-\alpha \sum_{i=1,2,3} x_i^2)$, where $\alpha = \frac{\gamma}{8\lambda\sigma^2}$ and $\gamma > 0$ controls the step size of the Gauss-Newton step.

3.2 The Image Likelihood

The image likelihood $p(I|I_i)$ is needed to evaluate the joint probability in Eq. (3). We expand $p(I|I_i)$ using the generative model in Fig. 1 and approximate the resulting integral using Laplace's method [19]:

$$p(I|I_i) = \int_{\Phi} p(I|\Phi, I_i)p(\Phi)d\Phi \approx p(I|\Phi_i^*, I_i)p(\Phi_i^*)\sqrt{(2\pi)^{3|\Omega|}/\det H}, \tag{10}$$

where Φ_i^* is defined in Eq. (4) and computed in the previous section. det denotes matrix determinant, H is the Hessian matrix with entries $-\frac{\partial^2 \log[p(I|\Phi,I_i)p(\Phi)]}{\partial v_j(x)\partial v_k(y)}|_{\Phi=\Phi_i^*}$, for all $x, y \in \Omega \subset \mathbb{R}^3$ and $j, k = \{1, 2, 3\}$, and $|\Omega|$ is the number of voxels.

We approximate the determinant of the Hessian by ignoring the second derivative terms and interactions between neighboring voxels, cf.[15]:

$$\det H \propto \prod_{x\in\Omega} \det\left(\nabla I_i(\exp(v)(x))(\nabla I_i(\exp(v)(x)))^T + \frac{9}{2}\lambda\sigma^2 \mathrm{Id}_{3\times3}\right), \tag{11}$$

where $\nabla I_i(\exp(v)(x))$ is the 3×1 gradient of the warped training image I_i and $\mathrm{Id}_{3\times3}$ is the 3×3 identity matrix.

3.3 Segmentation via EM

With our model instantiation, the solution of Eq. (5) cannot be found in closed form, since a mixture of factorized distributions is not factorized. Yet, an efficient solution to this MAP formulation can be obtained via Expectation Maximization (EM). The derivation of the EM algorithm is straightforward. Here, we present a summary. The E-step updates the weights associated with each training image:

$$m_i^{(n)} \propto p(I|I_i) \prod_{x\in\Omega} \pi_{L_i}(\hat{L}^{(n-1)}(x); \Phi_i^*(x)), \tag{12}$$

where $\hat{L}^{(n-1)}(x)$ is the segmentation estimate of the test image from the previous iteration and the weights sum to 1, $\sum_i m_i^{(n)} = 1$. The M-step updates the segmentation estimate through the following maximization:

$$\hat{L}^{(n)}(x) = \underset{L(x)}{\operatorname{argmax}} \sum_{i=1}^{N} m_i^{(n)} \log\left(\pi_{L_i}(L(x); \Phi_i^*(x))\right). \tag{13}$$

The M-step in Eq. (13) performs an independent optimization at each voxel $x \in \Omega$. Each of these optimizations simply entails determining the mode of a length $\mathcal{L}$ vector, where $\mathcal{L}$ is the number of labels. The EM algorithm is initialized with $m_i^{(1)} \propto p(I|I_i)$ and iterates between Equations (13) and (12), until convergence.

4 Experiments

We validate SNIP with 39 T1-weighted brain MRI scans of dimensions 256×256×256, 1mm isotropic. Each MRI was manually delineated by an expert anatomist into left and right White Matter (WM), Cerebral Cortex (CT), Lateral Ventricle (LV), Hippocampus (HP), Thalamus (TH), Caudate (CA), Putamen (PU), Pallidum (PA) and Amygdala (AM). We use volume overlap with manual labels, as measured by the Dice score [20], to quantify segmentation quality. The Dice score ranges from 0 to 1, with higher values indicating improved segmentation.

4.1 Setting Parameters through Training

SNIP has three independent parameters: (a) the image intensity variance σ^2 in Eq. (6), (b) the warp stiffness parameter λ in Eq. (8), and (c) the step size γ in the registration algorithm in Section 3.1. In particular, the registration component of SNIP is completely determined by γ and $\alpha = \frac{\gamma}{8\lambda\sigma^2}$, while the segmentation component is determined by σ^2 and λ.

Nine subjects were used to determine the optimal values of these parameters. First, 20 random pairs of these nine subjects were registered for a range of values of γ and α. Registration quality was assessed by the amount of pairwise label overlap and used to select the optimal (γ^*, α^*) pair.

We used the optimal (γ^*, α^*) pair to register all 72 ordered pairs of the 9 training subjects. We performed nine leave-one-out segmentations using these alignments with different pairs of σ^2 and λ that satisfy the relationship $\lambda\sigma^2 = \frac{\gamma^*}{8\lambda^*}$. The pair that yielded the best segmentation results was deemed optimal and used in validation on the remaining 30 subjects.

4.2 Benchmarks

First, we consider our implementation of the Label Fusion algorithm [12]. We use the pairwise registrations obtained with (γ^*, α^*) to transfer the labels to the training subject via nearest-neighbor interpolation. Segmentation is then computed through majority voting at each voxel. In the second benchmark, we use the label probability maps, where each training image voxel has a length-$\mathcal{L}$ vector, with one for the entry corresponding to the manual label, and zero otherwise. Segmentation for each voxel is determined to be the label corresponding to the mode of the label probability obtained by averaging the warped label probability maps, computed using the pairwise registrations and trilinear interpolation. We call this method Probabilistic Label Fusion.

4.3 Results

We report results for the 30 subjects not included in the group used for setting the algorithm parameters γ, σ, α. For each test subject, we treat the remaining subjects as training data. We note that the results from the two hemispheres are very similar and report results averaged across two hemispheres.

Fig. 2 shows box-plots of Dice scores for the two benchmarks and SNIP. These results indicate that SNIP outperforms the two benchmarks in all structures,

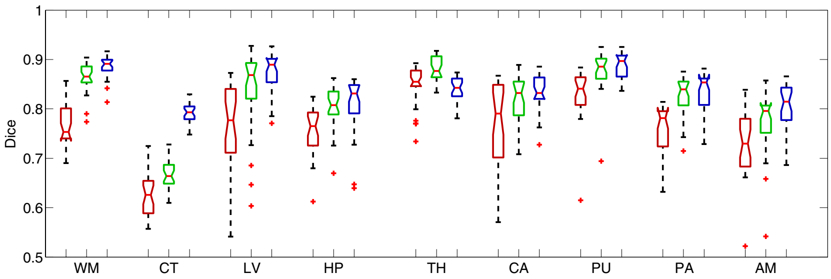

Fig. 2. Boxplots of Dice scores for Label Fusion (red), Probabilistic Label Fusion (green) and SNIP (blue). Medians are indicated by horizontal bars. Boxes indicate the lower and upper quartiles and vertical lines extend to 1.5 inter-quartile spacing.

Table 1. Comparison with FreeSurfer [1] and FreeSurfer Atlas Renormalization [2]

	HP	TH	CA	PU	PA	AM
SNIP	0.81 ± 0.07	0.84 ± 0.03	0.84 ± 0.04	0.89 ± 0.03	0.83 ± 0.04	0.80 ± 0.05
FS [1]	0.79 ± 0.09	0.88 ± 0.02	0.79 ± 0.10	0.81 ± 0.07	0.71 ± 0.09	0.71 ± 0.12
FSAR [2]	0.82 ± 0.04	0.88 ± 0.02	0.84 ± 0.05	0.85 ± 0.04	0.76 ± 0.06	0.75 ± 0.07

except the thalamus. The improvement is particularly significant in the White Matter, Cortex, Lateral Ventricle and Hippocampus. Between the two benchmarks, the performance of Probabilistic Label Fusion is consistently higher than that of Label Fusion. We note, however, that the results we report for SNIP are in the same ball-park as the ones reported for Label Fusion [12], and thus higher than what we achieve with our Label Fusion implementation. This might be due to differences in the data and/or registration algorithm. Specifically, normalized mutual information (NMI) was used as the registration cost function in [12]. Entropy-based measures such as NMI are known to yield more robust alignment results. We leave a careful analysis of this issue and an extension of SNIP that utilizes entropy-based similarity measures to future work.

Segmentation results for six subcortical structures were reported by two other state-of-the art atlas-based segmentation methods: FreeSurfer (FS) [1] and the FreeSurfer Atlas Renormalization (FSAR) technique [2]. Table 1 lists the average and s.t.d. of the dice scores reported in [1,2]. These results suggest that SNIP's performance is better for 3 ROIs (PU, PA, AM), equivalent for two ROIs (CA, HP) and worse for one ROI (TH).

The computational complexity of SNIP grows linearly with the number of training subjects. With the 39 training images we tested on, the segmentation procedure of each test subject took about 30 hours of CPU time on a modern computer. This run-time can be significantly reduced by solving the registrations in parallel. In comparison, Freesurfer took 10 hours and our Label Fusion implementation took 24 hours. Managing large training datasets within the SNIP framework is an important open question that we leave to future research.

5 Conclusion

This paper presents SNIP, a novel segmentation framework that adopts a non-parametric Bayesian approach to segmentation. By leveraging fast and robust nonrigid registration algorithms, SNIP exploits the entire training set, rather than a summary of it. In contrast to Label Fusion [12], similarities between the test image and training images play a central role in the segmentation. Our experiments indicate that SNIP promises to improve the performance of Label Fusion and compares favorably against other state-of-the-art atlas based segmentation methods in several regions of interest. One particularly promising future direction is to incorporate an entropy-based similarity measure into the computation of the image likelihood and pairwise registrations.

Acknowledgments. Support for this research is provided in part by: NAMIC (NIH NIBIB NAMIC U54-EB005149), the NAC (NIH NCRR NAC P41-RR13218), the mBIRN (NIH NCRR mBIRN U24-RR021382), the NIH NINDS R01-NS051826 grant, the NSF CAREER 0642971 grant, NCRR (P41-RR14075, R01 RR16594-01A1), the NIBIB (R01 EB001550, R01EB006758), the NINDS (R01 NS052585-01), the MIND Institute, and the Autism & Dyslexia Project funded by the Ellison Medical Foundation. B.T. Thomas Yeo is funded by the A*STAR,Singapore.

References

1. Fischl, B., Salat, D., van der Kouwe, A., Makris, N., Segonne, F., Quinn, B., Dale, A.: Sequence-independent segmentation of magnetic resonance images. Neuroimage 23, 69–84 (2004)
2. Han, X., Fischl, B.: Atlas renormalization for improved brain MR image segmentation across scanner platforms. IEEE TMI 26(4), 479–486 (2007)
3. Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J.: A probabilistic atlas of the human brain: Theory and rationale for its development: The international consortium for brain mapping (ICBM). Neuroimage 2(2), 89–101 (1995)
4. Pohl, K., Fisher, J., Grimson, W., Kikinis, R., Wells, W.: A bayesian model for joint segmentation and registration. Neuroimage 31, 228–239 (2006)
5. Yeo, B., Sabuncu, M., Desikan, R., Fischl, B., Golland, P.: Effects of registration regularization and atlas sharpness on segmentation accuracy. Medical Image Analysis 12(5), 603–615 (2008)
6. Guimond, A., Meunier, F., Thirion, J.: Average brain models: A convergence study. Computer Vision and Image Understanding 77(2), 192–210 (2000)
7. Joshi, S., Davis, B., Jomier, M., Gerig, G.: Unbiased diffeomorphism atlas construction for computational anatomy. Neuroimage 23, 151–160 (2004)
8. Ashburner, J., Friston, K.: Unified segmentation. Neuroimage 26, 839–851 (2005)
9. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Symmetric log-domain diffeomorphic registration: A demons-based approach. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part I. LNCS, vol. 5241, pp. 754–761. Springer, Heidelberg (2008)
10. Sabuncu, M., Yeo, B., Vercauteren, T., Leemput, K.V., Golland, P.: Asymmetric image template registration. In: Yang, G.-Z., et al. (eds.) MICCAI 2009, Part I. LNCS, vol. 5761, pp. 565–573. Springer, Heidelberg (2009)

11. Aljabar, P., Heckemann, R.A., Hammers, A., Hajnal, J.V., Rueckert, D.: Classifier selection strategies for label fusion using large atlas databases. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) MICCAI 2007, Part I. LNCS, vol. 4791, pp. 523–531. Springer, Heidelberg (2007)
12. Heckemann, R., Hajnal, J., Aljabar, P., Rueckert, D., Hammers, A.: Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. Neuroimage 33(1), 115–126 (2006)
13. Isgum, I., Staring, M., Rutten, A., Prokop, M., Viergever, M., van Ginneken, B.: Multi-atlas-based segmentation with local decision fusion-application to cardiac and aortic segmentation in CT scans. TM (in press, 2009)
14. Allassonniere, S., Kuhn, E., Trouve, A.: Construction of Bayesian deformable models via stochastic approximation algorithm: a convergence study. arXiv.org (2009)
15. Van Leemput, K.: Probabilistic brain atlas encoding using bayesian inference. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4190, pp. 704–711. Springer, Heidelberg (2006)
16. Davis, B., Fletcher, P., Bullitt, E., Joshi, S.: Population shape regression from random design data. In: Proc. of ICCV, pp. 1–7 (2007)
17. Warfield, S., Zou, K., Wells, W.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for validation of image segmentation. TMI 23(7), 903–921 (2004)
18. Arsigny, V., Commowick, O., Pennec, X., Ayache, N.: A log-euclidean framework for statistics on diffeomorphisms. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4190, pp. 924–931. Springer, Heidelberg (2006)
19. MacKay, D.: Information Theory, Pattern Recognition and Neural Networks (2003)
20. Dice, L.: Measures of the amount of ecologic association between species. Ecology 26(3), 297–302 (1945)